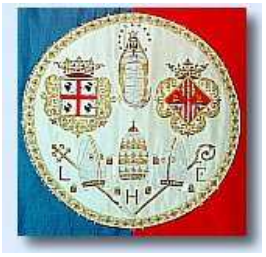

Università degli Studi di Cagliari

DOTTORATO DI RICERCA

Ciclo: **XXIII**

Titolo Tesi:

Reaction-Diffusion Processes on Complex Networks

Settore scientifico disciplinare di afferenza: **FIS/02**

Presentata da: **Nicola Perra**

Coordinatore Dottorato: **Prof. Gianluca Usai**

Tutor: **Prof. Gianni Mula**

Esame finale anno accademico 2009-2010

Reviewer: Prof. Luciano Pietronero

Day of the defense: 11/01/2011:

Signature from head of PhD committee:

Abstract

The availability of an unprecedented amount of large-scale data sets has allowed to discover complex features in many real-world networks. This boosted the attention on complex networks in different disciplines: Mathematics, Medicine, Biology, Social Sciences, Computer Sciences and Physics. Recently, a lot of attention has been devoted to the study of dynamical processes occurring on complex networks.
In this work we focused on the general framework of Reaction-Diffusion models on complex topologies. Using this approach we discussed two different problems.
The first one is based on the evaluation of the importance or centrality of nodes. In heterogeneous networks not all the nodes are the same. To sort out the differences is a relevant problem in data retrieval, biology and in general infrastructure management. The relative importance of units is not just a local feature. The centrality of a node is in fact related to the importance of the nodes that are connected to it and so on. Therefore we have a diffusion process, the diffusion of importance, which is encoded in the spectral properties of several kinds of matrices. Spectral centrality measures are accordingly defined and new results and the interpretations of these measures on directed, undirected and real networks are presented.
The second problem discussed within the same framework is the epidemic spreading in homogeneous and heterogeneous networks. This is an extremely relevant problem for our society as demonstrated by the last H1N1 pandemic in 2009. Complex networks analysis is crucial to get enough insight on the epidemic processes to suggest efficient interventions policies and make forecasts. We introduced the general theory of epidemic spreading on networks. We presented new single population models in order to deal with the effects of social disruption due to the epidemic diffusion itself. We discussed the framework of metapopulation network models in which each population is considered as a node of a network. The populations are coupled by diffusion of individuals. Markovian diffusion is first considered and all the known results are reproduced and derived. A more realistic protocol considering origin-destination matrices is introduced and analytically solved.
In the last chapter we show how these models can be used in order to build a realist data driven model, GLEaM, which is a powerful tool to make global epidemic forecasts. The use of this model during the recent H1N1 pandemic is described and all the new methods and results obtained are discussed in detail.

To Whom It May Concern

Contents

Preface

The work presented in this dissertation was mainly carried out at the School of Informatics and Computing at Indiana University in the Center for Complex Networks and Systems Research and at the Physics Department, University of Cagliari in the period between January 2008 and December 2010. Part of this work has been also done at Institute for Scientific Interchange (ISI) in Turin. I thank these institutions for their kind hospitality.

Most of the work presented in this thesis has been done within different scientific collaborations, whose members I kindly acknowledge for giving me the possibility and the honor to work with them.

Chapter 3 is the result of a collaboration with Santo Fortunato. It is based on the following paper:

Spectral centrality measures in complex networks published in Phys. Rev. E in the 2008.

Chapter 4 is the result of a collaboration with Vinko Zlatić, Alessandro Chessa, Claudio Conti, Debora Donato and Guido Caldarelli. It is based on the following paper:

PageRank equation and locatization in the WWW published in EPL in the 2009.

Chapter 5 is the rusult of a collaboration with Duygu Balcan, Bruno Goncalves and Alessandro Vespignani. It is based on a paper in preparation.

Chapter 6 is the result of a collaboration with Yamir Moreno, Sandro Meloni and Alessandro Vespignani. It is based on a paper in preparation.

Chapter 7 is the result of a collaboration with Duygu Balcan, Hao Hu, Bruno Goncalves, Paolo Bajardi, Chiara Poletto, Jose J. Ramasco, Daniela Paolotti, Michele Tizzoni, Andrew C. Singer, Christos Chouaid, Wouter Van den Broeck, Vittoria Colizza and Alessandro Vespignani. It is based on the following papers:

Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility published in BMC medicine in the 2009,

Estimate of Novel Influenza A/H1N1 cases in Mexico at the early stage of the pandemic with a spatially structured epidemic model published in PLoS current Influenza in the 2009,

Modeling the critical care demand and antibiotics resources needed during the Fall 2009 wave of influenza A(H1N1) pandemic published in PLoS current Influenza in the 2009,

Modeling vaccination campaigns and the fall/winter 2009 activity of the new A(H1N1) influenza in the northern hemisphere published in Emerging Health Threats in the 2009.

Introduction and Motivations

More is different.
P.W. Anderson

In the last years we have witnessed an intense research activity on complex systems. However a universally accepted definition of them is still missing. So, what is a complex system? It is important first to stress that *complex* does not mean something merely *complicated*. Though computer, cars, houses and airplanes are made of a huge number of elements designed to have and perform different tasks, they are really merely complicated, since they are engineered systems put in place according to a precise blueprint. On the contrary complex systems are emergent phenomena: they are the spontaneous outcome of the interactions among the system constitutive units. They are self-organizing systems, there is not blueprint or global supervision. Their behavior cannot be described from simple extrapolations of the properties of their constitutive units. The study of each subpart of the system in isolation does not allow an understanding of the whole system or of its dynamics. Another important feature of complex systems is the presence of structures whose fluctuations are extended and repeated at all possible scales. Critical phenomena (1) where macroscopic rearrangements across the entire system are generated by the combined action of infinitesimal localized interactions, or fractal objects (2) in which we see the same level of complications independently of the resolution used to look at the system, as shown in Figure (1), are just few examples of this feature.

Many complex systems admit an abstract mathematical representation as a graph in which nodes or vertices are the units/elements of the system and links or edges represent a relation or interaction among those elements. If, after this abstraction has been made, the complex features are still present (3) we will refer to these complex systems

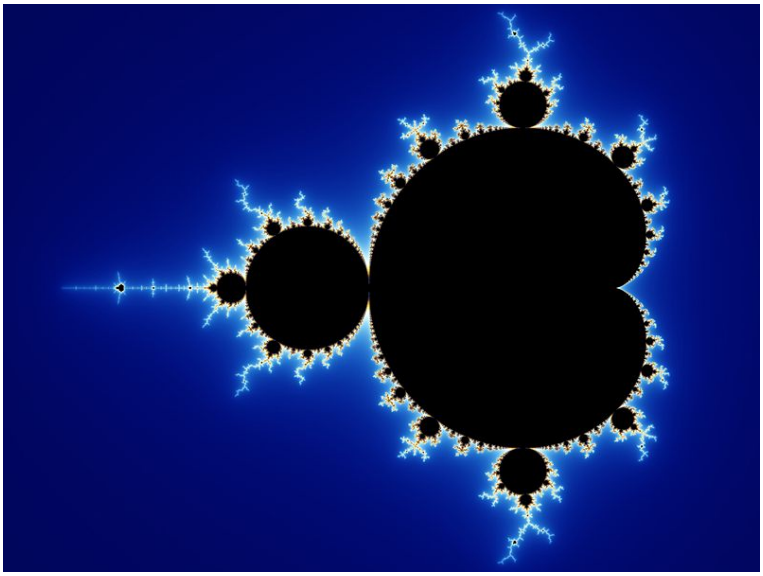

Figure 1: In this figure we show the famous representation of the Mandelbrot set $z_{n+1} = z_n^2 + c$. Created by Wolfgang Beyer with the program Ultra Fractal 3.

as complex networks. This level of abstraction applies to a huge number of systems across different scientific fields, from social interactions among individuals to biological interactions between proteins. The mathematical framework is based on Graph Theory (4; 5), one of the fundamental parts of discrete mathematics founded by Euler in 1736 with the famous solution of the Königsberg bridges problem.

In the last decade the availability of an unprecedented amount of data, due to the increase of computer power and communication networks, allowed a shift from the analysis of small graphs and the properties of single nodes or edges to the consideration of large-scale statistical properties of graphs. This change of perspective/scale is very important and we can compare it to the shift from the atomic scale and molecular physics to the physics of matter. Large complex networks arise in a vast number of natural and artificial systems: the brain, ecosystems, social systems (facebook, twitter, emails), transportation networks, power grids, the internet, the World Wide Web etc.. Millions or even billions of nodes can now be mapped. We know that such dramatic changes of scale require a change of the analytic/theoretical approaches and cause a new phenomenology to emerge. The present case is not an exception. *More is different* (6), the same elements assembled in large number can cause different macroscopic and dynamical be-

haviors, and holistic approaches have to be used. It has been proved in a huge number of cases that large-scale networks are characterized by complex topologies and a very heterogeneous structure. In particular the heterogeneity is shown in connectivity patterns that are statistically characterized by heavy-tailed distributions and large fluctuations. This means that many networks show a small but significant number of nodes, called *hubs*, usually more connected with respect to the average. Social networks frequently contain a few central individuals with many acquaintances, in the WWW there are just a few very popular websites with a huge number of links, in proteins there are just a few metabolites that take part in big fraction of all metabolic processes, there are just few airports with a huge number of connecting flights etc.. In these networks there is no characteristic scale: this is why they are often called *scale-free* networks. Another important feature of many networks is the so called *small-world* effect: the average distance between pairs of nodes is very small and typically increases only with the logarithm, or more slowly, of the number of vertices in the networks. This is completely different from what happens in regular lattices. This feature is well known in the sociological context as the *six degrees of separation* after the pioneering works of Milgram in 1967 (7). The study showed that a small number of acquaintances, six on average, is enough to create a connection between any two people chosen at random in the United States. Another important feature of many complex networks is the tendency to form groups of densely interconnected elements. Formally this is expressed by a high clustering coefficient.
Many efforts were done in order to model these peculiar properties. The simplest model has been proposed in the seminal works of Erdös and Rényi in the 60s (8; 9; 10). They used the simplest assumption, very far from reality: each pair of nodes is connected with a given probability p. Their model leads to homogeneous random networks. They show the small-world effect but have a very small clustering coefficient and a Poissonian connectivity distribution. In 1998 another important paper introduced a new mechanism that can provide both a small-world effect and a high clustering coefficient (11) and helped to understand the deep meaning of these two properties. However an important feature was still missing, since it is possible to prove that even in this case the connectivity is not skewed. In 1999 was published one of the most cited papers of the last decade in which Barabási and Albert (12) proposed a growth model with a heavy-tailed connectivity distribution, high clustering and small-world effect. They considered that in many real networks new edges are not connected randomly, but they tend to connect

to vertices which already have a large number of connections: *preferential attachment*. After their model many others have been proposed as variations of this mechanism and helped to understand the topology of networks.

The progress in the understanding the structure of complex networks generated a lot of attention to the implications that such topologies have on dynamical processes occurring on top of them. In this work we will focus on a class of these processes: the Reaction-Diffusion ones. Any phenomenon in which local quantities obey physical reaction diffusion equations can be modeled in this framework. Any processes in which particles that diffuse are subject to various reactions determined by the nature of the specific problem. It is a general class that allows us to ask several important questions. How can we asses the importance of different nodes or retrieve data on large information structures? The navigation and exploration of complex networks are clearly affected by the underlying connectivity. The heterogeneity of these systems implies, as we said before, that not all the nodes are the same. Not all the information in the website is relevant for us during a research on the WWW. Before Google revolution the search engines were not efficient. Their searching criteria did not take into account the role of the complex features of the network. If we had inserted in one the search engines the word in Yahoo! we would not have found its site as the first result of the research. In the 1998 Page and Brin (13) understood that the importance or relevance of a page is determined by the whole network. We can imagine that the importance of a page spreads to its neighbors and so on. An easy and efficient way to determine the importance of a node is the probability that a random walker surfing the network will visit that page. This process is a random walk, an old friend of physicist, more generally it can be seen as diffusion process on a complex network. Page and Brin used this idea to build the PageRank, an extraordinary simple measure to assess the importance of webpages. This represented the winning feature of the search engine Google and a revolution in the way we access information on the world wide web.

How pathogens spread in populations? The spread of the Black Death, the epidemic plague (14), in the 14^{th} century was mainly a spatial diffusion phenomenon. Accurate historical studies have shown that the disease propagation followed a simple pattern that can be described with good precision using classical continuous differential equations with a diffusive term. As shown during the SARS epidemic (15) or the recent H1N1 (16) the spread of such infectious diseases in modern populations is mainly due

Figure 2: In this figure we show the multiscale mobility network of short-range connections (commuting) and long-range connections (flights). Figure courtesy of Bruno Goncalves.

to commercial air travels. It is not anymore a simple diffusive phenomenon, the spatial structure of modern transportation networks must be considered. The mobility of people is characterized by different time scales. People commute everyday to go to work traveling small distances, on the other hand people travel thousand of miles using airplanes within a country or between countries. We have short-range and long-range connections that can be integrated in the same multiscale network as shown in Figure (2). This multiscale network is a techno-social system (17) with large-scale infrastructures whose dynamics and evolution are defined and driven by human behavior. It exhibits a dynamic self-organization and it is statistically very heterogeneous: it is a complex system. The understanding of such structure is crucial in any attempt to study a spreading of an infectious disease. People react, meeting each other in the bus, during the classes, in the workplaces and people diffuse, going to work, to classes, to holidays. We have interactions within a city and among cities through movements of individuals. Convenient models to describe these spatially structured interacting subpopulations are *metapopulation* models (18; 19; 20; 21). These models are used in many disciplines: population ecology, genetics and adaptive evolution, whenever the spatial structure of population plays a key role in the evolution of the system. Metapopulation models are based on the assumption that the system is characterized by highly fragmented structures in which population is localized in relatively isolated subpopulations connected by some flow of migration. Our society can be described within this paradigm. We have cities connected through individual mobility world-wide. This appealing possibility boosted the basic and theoretical research on these models. Subpopulations are thus being replaced by nodes populated by a certain number of particles, and, on the basis of empirical evidence, heterogeneous connectivity patterns and different diffusion mechanisms have been considered. Surprising analytical result have been found (22; 23) that allow us to understand the role of the complex structure on important quantities such the *global invasion threshold* for the infection of a macroscopic number of nodes/subpopulations and of its relation to the diffusion rate of individuals. Then the question is: have we enough knowledge of the mobility patterns in our society, of human behavior and of the dynamical processes on complex topology to model the global spreading of a infectious disease with sufficient precision and accuracy to make reliable forecasts?

The key point is the accurate knowledge of the mobility patterns. Somehow these data can be obtained but they refer to *stationary states* of social behavior on the physical

infrastructures. In the case of a catastrophic events, for example an extremely lethal pandemic, people may change their habits, they can decide to stay home, not to travel, not to go to work. Thus we can aspect that the techno-social system can be driven out of equilibrium. This is quite easy to imagine, but extremely hard to model. How can we model and consider the adaptive behavior of individuals or social disruption during a global outbreak?

All these challenging, difficult and intriguing questions motivate this work. Evaluating the importance of a webpage or modeling the spreading of an infectious disease are completely different problems that can be attacked within one general approach well grounded in Physics, namely that of Reaction-Diffusion processes. We present a walk, not random, into the complexity and in the holistic Physics perspective. In particular Chapter 1 is dedicated to the introduction of basic concepts of graph theory, real-world networks and basic models of them. In Chapter 2 the general framework of dynamical processes on complex networks is introduced, with particular focus on random walk and epidemic spreading. The effect of heterogeneity on the dynamics is considered and studied in detail. Chapter 3 is devoted to the application of diffusion processes in finding the importance of nodes in complex networks. Spectral centrality measures are formally introduced and studied in different models and real networks. The importance and role of diffusion is discussed and some new general results and interpretations are presented. In Chapter 4 we focus on one of these measure, the PageRank, proposing a new interpretation for this quantity in the form of several well known problems in physics, from the charge-distribution in an inhomogeneous medium to the wave-localization phenomena in quantum-physics. In this theoretical discussion a new method to compute the PageRank is proposed and discussed. In Chapter 5 behavioral changes and social disruption in a single population are considered. New models are formally presented and studied in order to solve one of the big issues in epidemiology. We cannot claim to have found *the* solution but the effects of different scenarios are presented and analyzed one by one. In Chapter 6 metapopulations models are introduced, general theory and results are presented and derived in details. New results considering more complicated diffusion processes with origin-destination are considered and analytically solved. In Chapter 7 a realistic model for the global spreading of infectious disease, GLEaM (21), is presented and its application to the recent H1N1 pandemic is fully discussed. In particular: a new method for the determination of crucial parameters of the disease based on Monte Carlo

likelihood analysis, an estimation of the initial number of cases in Mexico and a model of the demand of antibiotics resources are proposed and discussed in detail.

1

Networks and Graphs

Elegance is not a dispensable luxury but a quality
that decides between success and failure.
E. Dijkstra.

Contents

Networks are the basic ingredient of this work. In this Chapter we introduce the principal concepts of networks and graph theory. We define different type of networks and the basic measures for the characterization of them: degree, shortest path length, clustering coefficient, betweenness and all their statistical properties. We describe some real-world network and the classical models that have been proposed to model them.

1.1 Basic definitions in graph theory

Any system that can be abstractly imagined as a graph (a collection of *vertices* joined together by *edges*) is a network (24; 25). Such general definition applies to a wide spectrum of systems.

The mathematical description of these abstract objects is formalized in a vast mathematical field: Graph Theory. The first scientist to introduce the notion of graph was Leonard Euler in the famous work *Solutio problematis ad geometriam situs pertinentis* in 1736, where he solved the Königsberg bridges problem. Since this paper the field is constantly growing. For a deeper analysis of the subject we invite the reader to some classical books (4; 5; 26; 27; 28; 29; 30).

Undirected graphs

In an undirected graph, $G(V, E)$, V is a non-empty countable set of vertices/nodes and E is a non-empty countable set of unordered pairs of different vertices called edges/links. The edges (i, j) represent a connection between $i \to j$ and $j \to i$. The two vertices are called *adjacent, connected, neighbors or nearest neighbors*. The number of nodes, N, is the cardinality of V. The number of links, m, is the cardinality of E. The two parameters N and m are not independent. The maximum value of m is bounded by N:

$$m_{max} = \binom{N}{2} = \frac{N(N-1)}{2}, \tag{1.1}$$

that represent all possible pairs of vertices joined by edges. A graph in which $m = m_{max}$ is called *complete*.

A particular case of undirected graphs are trees. They are hierarchical graphs without cycles [1]. If each node has exactly one parent is easy to show that:

$$N = m + 1. \tag{1.2}$$

Direct graphs

In a directed graph or digraph, $D(V, E)$, the set of edges E is ordered. The connection (i, j) between i and j implies just $i \to j$. The reverse connection is not necessarily present.

[1] A cycle is a closed path that visits each node, apart from the end-vertices, only one

Weighted graphs

We can imagine to assign to nodes and links more properties. An example of such generalization is provided by weighted graphs. In this case the new degree of freedom is given by the intensity of the connections. Examples such as the frequency of social interactions (31), the traffic of data in internet routers (32), the air traffic (33) are just few instance in which the simple topology is not enough to characterize the system: it is crucial to take into account that some edges are more important than others. In this case for each connection (i, j) we have to assign its weight w_{ij}.

Bipartite graphs

An undirected graph is called bipartite if it is characterized by two independent sets V_1 and V_2 of different type of nodes. Formally we define these graphs as $G = (V_1 + V_2, E)$. Every connection is made between the two different set. Then for each link (i, j) we will have $i \in V_1$ e $j \in V_2$ or vice versa.

Subgraphs

A graph $G' = (V', E')$ is called subgraph of $G = (V, E)$ if $V' \subset V$ and $E' \subset E$. A clique is a complete subgraph of size $n < N$.

Adjacency matrix

A graph is defined by its *adjacency matrix* $\mathbf{A}$, defined such that

$$a_{ij} = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{if } i \not\to j \end{cases} \tag{1.3}$$

If it is not explicitly specified there are not self-loops, then for $\forall\ i,\ a_{ii} = 0$. For undirected graphs the adjacency matrix is symmetric: $a_{ij} = a_{ji}$. For directed graphs in general we have $a_{ij} \neq a_{ji}$. For weighted graphs each element of the matrix will be a real number, the weight, indicated as a^w_{ij}.

Path, distance and diameter

P_{i_0,i_n} is a *path* in the graph $G = (V, E)$. It connects the node i_0 to the node i_n and it is defined by $n+1$ nodes $V_p = \{i_0, i_1, ..., i_n\}$ and n links $E_p = \{(i_0, i_1), (i_1, i_2), ..., (i_{n-1}, i_n)\}$.

The length of the path is n. In the special case in which $i_0 = i_n$ we have a cycle. The path that minimize the length is called *geodetic*, *distance* or *shortest path*.
For two given nodes i e j the *distance* $l_{ij} = l_s$ between them is the path with the minimum number of links between them. Let us consider $\mathcal{P}_{ij}$ as a generic path between i and j, using the adjacency matrix we can write:

$$l_s = min \sum_{k,l \in \mathcal{P}_{ij}} a_{kl} = min \sum_{k,l \in \mathcal{P}_{ij}, a_{kl} \neq 0} \frac{1}{a_{kl}}. \tag{1.4}$$

These definitions can be generalized easily for direct and weighted graph.
The *diameter* D of a graph is the maximal distance among all the pairs of nodes.

Dimension

The *dimension* of a graph is the average of the (1.4) considering all the pairs: $\langle l_s \rangle$. It is possible that different pairs of nodes have the same l_s. We can define the probability $P_l(l_s)$ to find two nodes characterized by a distance l_s. Considering this probability we can write:

$$\langle l_s \rangle = l = \sum_l l_s P_l(l_s) \equiv \frac{2}{N(N-1)} \sum_{i<j} l_{ij}. \tag{1.5}$$

In the general case is not necessary to have a path between all the pairs of nodes. If it is the case, the distance between to nodes that have not a path among them is not well defined. For this reason in the (1.5) are considered just all the connected pairs. There is an alternative definition of average geodetic distance:

$$l^{-1} = \frac{2}{N(N-1)} \sum_{i<j} l_{ij}^{-1}. \tag{1.6}$$

In this definition all the pairs are considered. If there is not a path between two nodes the l_{ij} is set to infinity.

Components

The *component* C of a graph is a connected subgraph. Two components $C_1 = (V_1, E_1)$ and $C_2 = (V_2, E_2)$ are disconnected if there is not a path $P_{i,j}$ such that $i \in V_1$ and $j \in V_2$. The adjacency matrix of a network with more than one component can be always written in block diagonal form. The non-zero elements of the matrix are confined to square blocks along the diagonal of the matrix with all other elements equal to zero.

It is important to stress that in general to produce this form the node labels must be chosen correctly. The choice of labels is completely arbitrary, and has no effect on the structure of the graph itself. In such graphs it is possible to find a set of labels that will produce a matrix in block diagonal form. Some graphs are characterized by the presence of a *giant component* (GC) defined as a component whose size scale with the number of nodes, and diverges in the limit $N \to \infty$. The presence of a giant component implies then that a macroscopic fraction of the graph is connected.
In the case of direct graph the structure of components is more complex since the presence of a path between i and j does not imply the presence of a path between j and i. We can define components for a directed network as the strongly connected components, that are defined as the maximal subset of vertices such that there is a directed path in both directions between every pair in the subset. Every vertex in such component must belong to at least one cycle. Acyclic directed graph have no strongly connected components with more than one node. The component structure of a directed network can be decomposed into a giant weakly connected component, generally indicated (GWCC), that correspond to the giant component of the same graph in which the links are considered as undirected and a set of smaller disconnected components. The (GWCC) is composed of several parts, the giant strongly connected component (GSCC), formed considering the subset of nodes that have a directed path joining any pair of them, the giant in-component (GIN) formed by the nodes form which it is possible to reach the (GSCC) through a direct path, the giant out-component (GOUT) formed by the nodes that can be reached from the (GSCC) through a directed path and the tendrils containing nodes that cannot be reach or be reached by nodes into the (GSCC). Examples of those are the tube that connect the (GIN) and (GOUT) without pass through the (GSCC).

Clustering coefficient

The *clustering coefficient* C_i is defined as the fraction of pairs connected to a node that have a connection among them too. Let us consider a vertex i connected with three other nodes. Let us imagine that two of them are connected each other. In this case $C_i = 1/3$ because just 1 pair among 3 is actually connected. For a complete graph the clustering coefficient has is maximum value $C_i = 1 \ \ for \ \forall \ \ i \ \ \in G$. We can think this quantity as the fraction of links in the graph on respect to the maximum number between any i, j, k.

Using the adjacency matrix and the (1.1) we have:

$$C_i = \frac{2}{k_i(k_i-1)} \sum_{j,k} a_{ij} a_{ik} a_{jk} \tag{1.7}$$

The generalization of the clustering coefficient for directed graph is not easy. It is possible the definition of two coefficients for in and out degree.

1.2 Centrality measures

The adjacency matrix entirely define the structure of the graph. We can introduce a variety of measures able to capture features of the topology. One of the most crucial variety of measures are devoted to find which are the most important or central vertices in a network. These measures are called centrality measures. The most commonly used are the degree, closeness and betweenness centrality. Another very important centrality measures will be analyze in details in the Chapter 3.

Degree

The number of links connected to a generic node is called degree or connectivity. It is the easiest centrality measure and it is typically indicated as k. Considering the (1.3) it is easy to understand:

$$k_i = \sum_{j=1,n} a_{ij}. \tag{1.8}$$

For directed graphs instead we have to split this quantity in *in-degree* (incoming links) and *out-degree* (outgoing links)

$$k_i^{in} = \sum_{j=1,n} a_{ij}^T, \tag{1.9}$$

$$k_i^{out} = \sum_{j=1,n} a_{ij}. \tag{1.10}$$

For a weighted graph we define the *weighted degree*: k_i^w as:

$$k_i^w = \sum_{j=1,n} a_{ij}^w. \tag{1.11}$$

Usually this quantity is call *strength*. It has be proven that for different real network, strength and degree are related (34):

$$k_i^w \propto k_i^\eta. \tag{1.12}$$

Closeness centrality

The closeness centrality is defined as the average distance of a vertex to all the others:

$$g_i = \frac{1}{\sum_{j \neq i} l_{ij}}. \tag{1.13}$$

Of course, the nodes with a small shortest path distance to the other nodes have a large closeness centrality.

Betweenness centrality

The previous measures consider nodes which are topologically better connected to the rest of the graph. Another class of nodes are vertices that are crucial for connecting different regions of the network. In order to measure quantitatively the role of such nodes the concept of betweenness centrality has been introduced (35). Formally the *betweenness* of a vertex i is defined as the fraction of geodetic paths among any pair of vertices that pass through i:

$$b(i) = \sum_{\substack{j,l=1,n \\ i \neq j \neq l}} \frac{\mathcal{D}_{jl}(i)}{\mathcal{D}_{jl}}, \tag{1.14}$$

where $\mathcal{D}_{jl}$ is the total number of geodetic from j to l and $\mathcal{D}_{jl}(i)$ is the number of geodetic from j to l that goes through i. Unfortunately the calculation of this measure is computationally very consuming, order $\mathcal{O}(N^2E)$ that is prohibitive for large networks. Efficient algorithms has been proposed, for example in the Ref. (36) the complexity is reduced to an order $\mathcal{O}(NE)$ for unweighted networks.

1.3 Statistical properties

A statical characterization is needed for the study of the properties of the graph as a whole. In this section we will introduce statistical distribution of the quantities previously defined.

Degree distribution

The degree distribution $P(k)$ of a undirected graph is the probability to find a node with degree k. The average degree $\langle k \rangle$ is:

$$\langle k \rangle = \sum_k kP(k) \equiv \frac{2E}{N}. \tag{1.15}$$

A graph is called *sparse* if the average degree is very small on respect to the number of nodes: $\langle k \rangle \ll N$. In the case of directed graph we have two distributions $P(k_{in})$ for the in-degree and $P(k_{out})$ for the out-degree. It is easy to understand that:

$$\langle k_{in} \rangle = \sum_{k_{in}} k_{in} P(k_{in}) = \langle k_{out} \rangle = \sum_{k_{out}} k_{out} P(k_{out}) \equiv \frac{\langle k \rangle}{2}. \tag{1.16}$$

As we will discuss in details in the next sections for a wide range of degree distributions we have strong fluctuations on respect to the average value. It is important the study of the moments of the degree distribution:[1]

$$\langle k^n \rangle = \sum_k k^n P(k). \tag{1.17}$$

The moment of order two is also knows as the variance of the distribution. It plays an important role in dynamical process on network as we will see in the Chapter 2.

Joint probability

The probability $P(k, k')$ that a random link will connect two nodes of degree k and k' is called *joint probability*. The conditional probability $P(k|k')$ define the probability that given a node of degree k it will be connected with a node of degree k'. The probability to pick up randomly a node of degree k is $P(k)$. In the case we will extract a link the probability that we will find node of degree k is proportional to $kP(k)$ and we will call it $P^{end}(k)$. More precisely:

$$P^{end}(k) = \frac{kP(k)}{\sum_k kP(k)} = \frac{kP(k)}{\langle k \rangle}. \tag{1.18}$$

[1]The moment of order n of a generic distribution $g(x)$ is defined as : $\langle x^n \rangle = \sum_x x^n g(x) \overset{\Delta x \to 0}{\to} \int dx x^n g(x)$

We can now find the joint probability $P(k, k')$. This will be proportional to the number of links $E_{k,k'}$ between the vertex of degree k and k'. Considering the opportune normalization we will have:

$$P(k, k') = \frac{E_{k,k'}}{\sum_{k,k'} E_{k,k'}} = \frac{E_{k,k'}}{\langle k \rangle n}. \tag{1.19}$$

It is clear how:

$$\sum_{k'} P(k, k') = P^{end}(k) = \frac{kP(k)}{\langle k \rangle}. \tag{1.20}$$

Now since $P(A|B) = \frac{P(AB)}{P(B)}$ we will obtain:

$$P^{end}(k')P(k|k') = P(k, k'), \tag{1.21}$$

from which considering the (1.19) we will get:

$$P(k|k') = \frac{P(k, k')}{P^{end}(k')} = \frac{\langle k \rangle P(k, k')}{k'P(k')} = \frac{E_{k,k'}}{n_{k'}k'}. \tag{1.22}$$

The probability must be normalized:

$$\sum_{k} P(k|k') = 1. \tag{1.23}$$

The number of links from a vertex of degree k through a vertex of degree k' is equivalent in an undirected graph to the number of links from a vertex of degree k' through a vertex of degree k. This argument can be encoded in the so called detailed balance equation (37):

$$k'P(k|k')P(k') = kP(k'|k)P(k). \tag{1.24}$$

Correlation function

Many real networks such as the network of scientific collaboration show a correlation between the degree of a node and the degree of its neighbors i.e. nodes with high degree are preferentially connected with high degree nodes. This kind of correlation is called *assortative mixing*. In other cases the opposite situation has been found i.e. high degree nodes connected preferentially with low degree nodes. This kind of correlation is called *disassortative mixing* (38). Formally these correlations are measured considering the average degree of the nearest neighbors of a generic node i, $k_{nn,i}$:

$$\overline{k}_{nn}(k) = \frac{1}{N_k} \sum_{i} k_{nn,i} \delta_{k_i,k}, \tag{1.25}$$

where the sum runs over all nodes and N_k is the total number of nodes of degree k. Considering $P(k'|k)$ we can rewrite the (1.25) as:

$$\overline{k}_{nn}(k) = \sum_{k'} k' P(k'|k). \tag{1.26}$$

In the case in which there is not correlation we will have:

$$\frac{d(k_{nn})}{dk} = 0. \tag{1.27}$$

In the case in which $\overline{k}_{nn}(k)$ is an increasing function of k we will have an *assortative mixing*: vertices with high degree are more luckily connected with high degree nodes. In the case in which $\overline{k}_{nn}(k)$ is a decreasing function of k we will have a *disassortative mixing*: vertices with high degree are more luckily connected with low degree nodes. We can define an assortativity measure: the coefficient r. Considering the correlation function:

$$\langle kk' \rangle - \langle k \rangle \langle k' \rangle = \sum_{k,k'} kk' \left(P(k,k') - P(k)P(k') \right). \tag{1.28}$$

Normalizing this with the variance:

$$r = \frac{1}{\sigma^2} \sum_{k,k'} kk' \left(P(k,k') - P(k)P(k') \right). \tag{1.29}$$

In the case in which $P(k,k') = P(k)P(k')$ no correlation will be present and $r = 0$. Instead if $r > 0$ we will have an assortative mixing and if $r < 0$ disassortative mixing. We can generalized these concepts for weighted graphs. For these cases the correlation usually studied is the average $\langle a^w_{ij} \rangle$ with the product $k_i k_j$. In same cases we have:

$$\langle a^w_{ij} \rangle \propto (k_i k_j)^\theta. \tag{1.30}$$

For air transportation network as shown in (33) we will have $\theta = 0.5 \pm 0.1$. Considering the (1.12) we can find the relation between the exponent θ and η:

$$k^w_i \propto k^\eta_i \simeq \langle a^w_{ij} \rangle k_i \simeq k_i (k_i k_j)^\theta, \tag{1.31}$$

that reads: $\eta = 1 + \theta$.

Betweenness distribution

Using the (1.14) we can define the betweenness distribution $P_b(b)$. This gives the probability that a node has betweenness b. We can evaluate the n^{th} moments of the distribution:

$$\langle b^n \rangle = \sum_b b^n P_b(b) \equiv \frac{1}{N} \sum_i b_i^n. \tag{1.32}$$

Distance distribution

In the literature there are two principal definitions of distance distribution:

- the probability distribution to find two nodes at distance l, $P_l(l)$
- the average number of nodes characterized by a distance smaller or equal to l: $M(l) = N \sum_{l'=0}^{l} P_l(l')$. For $l = 0$ we get the starting node so $M(0) = 1$. For $l = 1$ we have all the neighbors so that $M(1) = k + 1$. This number is dramatically different for different type of networks. The study of this quantity is then crucial to understand the structure of the considered graph.

1.4 Real networks

In this section we will give to the reader a quick overview of networks in real world. We will focus in particular on social, technological and biological networks.

1.4.1 Social networks

Social Networks represent the individuals as nodes and the social interactions among them (friendship, sexual relations, belonging to the same group of work) as links. This kind of networks has been studied since 1934 in the works of Moreno (39) and are extremely important not just for social sciences but even for a wide variety of processes from the spreading of infectious diseases to the emergence of consensus and knowledge diffusion. The historical problem related to these networks was the difficulty to get reliable information of a sufficient large number of people. Fortunately the recent explosion of online social interactions has made available data sets of unprecedented size. E-mail exchanges (40; 41), habits and shared interest inferred from web visits and professional

communities such as collaboration networks of film actors (11; 12; 42) or company directors networks (43) or co-authorship among scientist (44; 45; 46) are classical examples of these type of networks.

1.4.2 Technological networks

Technological networks are human-built networks design to accomplish the distribution of some resource: water, electricity, gas etc.. A classical examples are: the networks of power grids both high or low voltage (11; 47), the networks of inter-urban streets (48), internet (32; 49; 50) and the airport networks (33; 51). This last system can be represented as a weighted graph where nodes are the airports, links the air connections and weights the flow of passengers. For more details we invite the reader to the Chapter 7 were a complete description of such network is presented. Another important technological network often classified as an information networks is the *WolrdWideWeb*. It is the most famous virtual network were nodes are web pages and links are the hyper-links (direct links) between them. The Web growth is extremely rapid and unregulated has led to a huge complex network. The structure of it is very difficult to study and for many years experiments has been done in order to get information about it (52; 53).

1.4.3 Biological networks

Biological networks completely pervade the biological world. From microscopic realm of biological chemistry, genetics, proteomics and large scale food webs. An important example are protein interaction networks (PIN) of various organisms where nodes represent proteins and edges connecting pairs of interacting proteins (54; 55). Three different scales of process are usually considered. The microscopic scale such PIN networks in which the main point is to understand the biological significance of the topology of this networks (56). At a larger scale biological networks can describe interactions between animals and even humans (57). At the very larger scale we find the networks describing the food web of entire ecosystems (58).

1.4.4 Small-World phenomenon

If we consider a regular lattice the average shortest path length $< l >$ follow the well known scaling $< l > \sim N^{1/2}$. The small-world property refers to networks in which $< l >$

scales logarithmically or slowly with the number of vertices. In many cases we have not data on the same network at different sizes then the small-world properties refers to the behavior of the quantity $M(l)$ defined as the average number of nodes within a distance less or equal to l from any given node (for more details we invite the reader to the section 1.3). For regular lattice this quantity increase as a power law with the distance l in small-world networks this quantity follows an exponential or faster increase. The really interesting fact is that this property is typical of many real networks more than just a mathematical obscure particular case. Also known as *six degrees of separation* phenomenon the small-world effect has become famous in the sociological context by Milgram in the 1967 (7). In his experiment he show how with on average six number of acquaintances is possible to connect any two people chosen at random in the United States. The same property has been found in a wide variety of networks in particular in technological networks.

1.4.5 Heterogeneity and heavy tails

Networks can generally divided into two class according to functional from of the statistical distributions of degree, betweenness and weighted quantities: homogeneous and heterogeneous. In the first class we have a fast decay of these quantities such Gaussian or Poisson distributions. In the second class instead we face with heavy-tailed distributions. What this means or imply? Let us consider for example the WWW, some page such *Google* become extremely popular and are linked by a huge number of other pages while in general most pages are almost unknown. The same thing happen in the airport networks: millions of person flight through Fiumicino in Rome but just few passenger flight through Olbia. These kind of networks are then characterized by *hubs*: node with a connectivity orders of magnitude lager then the average values. It turns out that many real-world networks has this kind of degree distribution (12). Heavy tails can be approximated by power-law decay $P(k) \sim k^{-\alpha}$. That means that vertices with a degree much more higher than the average are found with a no zero probability, so there is not a characteristic scale like in the cases homogeneous networks with a bell-shaped distributions and fast decaying tails. Let us consider explicitly the average degree:

$$< k >= \int dk k P(k). \tag{1.33}$$

If we consider a power law distribution with $2 < \alpha \leq 3$ the average is well defined and finite. If we consider higher moments like for example the typical second order moment:

$$< k^2 >= \int dk k^2 P(k). \tag{1.34}$$

we will realize that in the asymptotic limit of $k \to \infty$ the second moment is divergent: so fluctuations are unbounded and depend on the system size. This is due to the absence of any intrinsic scale, we observe then a *scale-free* network. If we consider $k^{-\alpha}$ and we consider a rescaling with some constant $k \to \lambda k$ the distribution will be $\lambda^{-\alpha} k^{-\alpha}$ so the shape is the same at any scale we consider. This is an important properties of such distribution that is called *self-similarity*.
One possible argument against the presence of these feature in real networks could be that heavy-tail truncation is the natural effect of the upper limit of the distributions. This is a generally true consideration, but if the fluctuations are orders of magnitude (three or more) than expected values we can consider the presence of a scale-free behavior not just due to the finite size of the system.

1.5 Graph models

After the brief description of some real networks and their properties in this section we will introduce basic graph models that have been proposed to describe them.

1.5.1 Erdös-Rényi (ER) model

One important topic on graph theory has been the study of random graph. The main contribution are due to Paul Erdös and Alfréd Rényi (8; 9; 10). In their first work they defined a random graph of N vertices and m links selected random among the $N(N-1)/2$. There are in total $C^n_{n(n-1)/2}$ graph. They can appear with the same probability and they form the ensemble of graph characterized by this rule. An equivalent definition of random graph is given by the binomial model. Starting from N vertices for each pair of nodes a link is formed with probability p. The number of links is then a random variable with average value $m = pN(N-1)/2$. Analytically quite often the asymptotic limit $N \to \infty$ is studied. According to Erdös and Rényi any graph is characterized by some properties Q in the case in which the probability that this properties will appear goes to the unity for $N \to \infty$. For different properties a critical

probability $p_c(N)$ exist. In the case in which $p(N)$ grows slowly of $p_c(N)$ the properties Q will not be shown and vice versa. We can write then:

$$\lim_{n\to\infty} P_{n,p(n)}(Q) = \begin{cases} 0 & \text{if } \frac{p(n)}{p_c(n)} \to 0 \\ 1 & \text{if } \frac{p(n)}{p_c(n)} \to \infty \end{cases} . \tag{1.35}$$

Degree distribution

In a random graph characterized by a probability of connection p the degree k_i follows the binomial distribution:

$$P(k_i = k) = C^k_{N-1} p^k (1-p)^{N-1-k}, \tag{1.36}$$

because the probability to be connected with k nodes is p^k, the probability to have not any other connections is $(1-p)^{N-1-k}$ and there are C^k_{N-1} possibles ways to select k vertices. The expectation value of the number of nodes of degree k, X_k, is:

$$E(X_k) = NP(k_i = k) = \lambda_k. \tag{1.37}$$

The distribution of X_k in the limit for $N \to \infty$ will be the Poisson distribution:

$$P(X_k = r) = e^{\lambda_k} \frac{\lambda_k^r}{r!}. \tag{1.38}$$

For big values of r the (1.38) goes rapidly to zero. A characteristic scale is then defined for $X_k = nP(k_i = k)$ with standard deviation $\sigma_k = \sqrt{\lambda_k}$.

Diameter

Random graph are characterized by very small diameter. They are typical *small-world* graphs. It is easy to show that the diameter l is proportional to $ln(n)/ln(\langle k \rangle)$ (59). An important feature of many real world networks is then reproduced by this model.

Clustering coefficient

Random graph are characterized by very small clustering coefficient. Given a node i the probability that two of its neighbors are connected is equal to the probability that any other two nodes will be connected then :

$$C_{rand} = p = \frac{\langle k \rangle}{N}. \tag{1.39}$$

This implies that the ratio $C_{rand}/\langle k \rangle$ goes like N^{-1}.

1.5.2 Watts-Strogatz (WS) model

As shown in the previous section, real networks are characterized by the small-word effect (on contrary to regular lattice and like random graphs) and relativity big values of clustering coefficient (on contrary to random graph and like regular lattices). Just from this simple observation we can conclude that real networks are not neither regular lattice or random graphs. The famous WS model (11) has been the first attempt to interpolate between these to limits. The model goes like that: given N nodes in a ring each node is connected to $k/2$ to the left and $k/2$ nodes to the right. The starting network is a sparse but connected graph provide that $n \gg k \gg ln(n) \gg 1$. Let us consider now p a rewire probability. Every link in the starting network is reassign random with probability p. With this process $pNk/2$ links will be change on average. Node that before were far away each other are now closer, thanks to the present of these short cuts before not allow by construction. For $p = 0$ we have a complete regular ring. For $p = 1$ a random graph. For any other value of p we will have a situation between this to limit. It is extremely interesting to study the behavior of the dimension of the graph and of the clustering coefficient as a function of p. In the two limit cases:

1. $l(0) \simeq \frac{n}{2k} \geq 1$, $C(0) \simeq \frac{3}{4}$,
2. $l(1) \simeq \frac{ln(n)}{ln(k)}$, $C(1) \simeq \frac{k}{n}$.

Interestedly enough as shown in Figure (1.1) as p increase the distance get reduced a lot. Instead the average clustering is almost constant. The two quantity change with a complete different slope with p. There are then regions in which we see a small-world effect and a value of clustering bigger than the random case. This kind of graph reproduce an important feature observed in many real network.

Degree distribution

In this model for $p = 0$ every node has the same degree:

$$P(k') = \delta(k' - k). \tag{1.40}$$

for $p > 0$ the game is a bit more complicate. As shown in Ref. (60) the degree k_i can be written as $k_i = k/2 + c_i$ where c_i can be divided in two part $c_i^1 \leq k/2$ considering that with probability $1 - p$ links are not changed and $c_i^2 = c_i - c_i^1$ considering that if a link

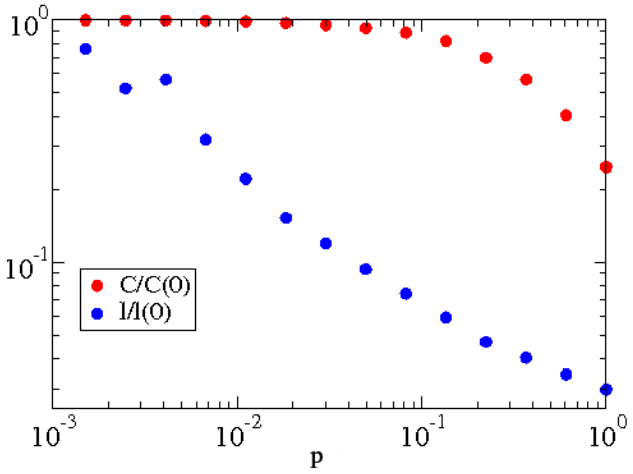

Figure 1.1: Average dimension of the graph l and average clustering coefficient for different values of p. $N = 10^3$ nodes and $k = 5$.

is rewired it is assigned to a node i with probability$1/n$. The distribution for c_i^1 and c_i^2 are then:

$$P_1(c_i^1) = C_{k/2}^{c_i^1}(1-p)^{c_i^1}p^{k/2-c_i^1}, \tag{1.41}$$

and:

$$P_2(c_i^2) = C_{pnk/2}^{c_i^2}\left(\frac{1}{n}\right)^{c_i^1}\left(1-\frac{1}{n}\right)^{pnk/2-c_i^1} \simeq \frac{(pk/2)^{c_i^2}}{c_i^2!}e^{-pk/2}, \tag{1.42}$$

that holds in the limit for large N. Combining the (1.42) with the (1.41) we get:

$$P(k') = \sum_{n=0}^{f(k,k')} C_{k/2}^{n}(1-p)^n p^{k/2-n}\frac{(pk/2)^{k'-k/2-n}}{(k'-k/2-n)!}e^{-pk/2}, \tag{1.43}$$

for $k' \geq k/2$, where $f(k',k) = min(k'-k/2, k/2)$. The shape of this distribution is similar to the distribution of a random graph. It is peaked for $\langle k' \rangle = k$ and has an exponential decay for big values of k'. The topology is then quite homogeneous.
This model is very important because describe a mechanism able to create a small-world with a high clustering coefficient. The degree distribution is instead far from the heavy-tailed distribution observed in real networks.

1.5.3 Barabási-Albert (BA) model

In this model, one of the most cited work in the literature in the last 10 years (61), gives a simple and reasonable mechanism able to produce scale-free graphs with small-world phenomena and high clustering on respect to a random graph.
The nodes here are progressively added. We have a growth of the network. In the previous models links were drown randomly. In this case we have different rules. If we think for example to the WWW the assumption that web pages are connected random appears to be completely out of reason. It is instead more reasonable to think that new pages will be connected to popular older pages. These changes are the basic of the BA model. The graph is build following these rules:

1. we start from a small number of core nodes n_0, at each time step a new vertex is added to the graph and it will be linked with $m < n_0$ other nodes already present

2. the probability π that the new node will be connected to the node i is function of the degree k_i: *preferential attachment*

$$\pi(k_i) = \frac{k_i}{\sum_j k_j}. \tag{1.44}$$

Numerical simulation shows that the degree distribution of such networks will be:

$$P(k) \simeq k^{-\gamma_{BA}} \qquad with \quad \gamma_{BA} = 3. \tag{1.45}$$

So, a scale-free distribution of degree.
There are different methods to evaluate analytically the degree distribution: the *continuum theory*, (12), *master-equation* approach, (62), *rate-equation* approach, (63). We will consider just the first approach here. The probability that at each time step the node i will increase its degree is $\pi(k_i)$. Let us now consider k_i as our real and continuous variable:

$$\frac{\partial k_i}{\partial t} = m\pi(k_i) = m\frac{k_i}{\sum_{j=1}^{n-1} k_j}. \tag{1.46}$$

Considering that $\sum_j k_j = 2mt - m$ we get:

$$\frac{\partial k_i}{\partial t} = \frac{k_i}{2t}. \tag{1.47}$$

The solution of this is

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\beta} \quad with \quad \beta = \frac{1}{2}. \tag{1.48}$$

The equation (1.48) requires that the degree of each node follow a power law with exponent β. We can now write down the probability that a node will have a degree $k_i(t)$ smaller of some k

$$P[k_i(t) < k] = P\left(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}\right). \tag{1.49}$$

Since we are adding nodes at regular rate:

$$P(t_i) = \frac{1}{n_0 + t}. \tag{1.50}$$

Using the (1.50) in the (1.49) we get:

$$P\left(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}\right) = 1 - \frac{m^{1/\beta}t}{k^{1/\beta}(t + n_0)}. \tag{1.51}$$

We can write the distribution $P(k)$ using

$$P(k) = \frac{\partial P[k_i(t) < k]}{\partial k} = \frac{2m^{1/\beta}t}{n_0 + t}\frac{1}{k^{1/\beta+1}}, \tag{1.52}$$

for $t \to \infty$

$$P(k) \sim 2m^{1/\beta}k^{-\gamma} \quad con \quad \gamma = \frac{1}{\beta} + 1 = 3, \tag{1.53}$$

that is in perfect agreement with numerical simulations.

Diameter

It is easy to show how the diameter of a BA graph is smaller of the relative measure of a random graph. The scale-free distribution is then characterized by a small-world effect. The diameter of a BA graph is

$$l \sim \frac{\ln n}{\ln \ln(n)}, \tag{1.54}$$

as shown in (64)

Degree correlations

As shows in Ref. (65) in the BA model we have degree-degree correlations. Let us consider all the pairs of degree k and l linked each other. Without lack of generality let us assume that the node with degree k was added after the other one. Then we have $k < l$ by construction (1.48). Let us consider the case $m = 1$. And let us define $N_{kl}(t)$ as the number of pairs of linked nodes of degree k and l. We have:

$$\frac{dN_{kl}}{dt} = \frac{(k-1)N_{k-1,l} - kN_{kl}}{\sum_k kN(k)} + \frac{(l-1)N_{k,l-1} - lN_{kl}}{\sum_k kN(k)} + (l-1)N_{l-1}\delta_{k1}. \tag{1.55}$$

We can rewrite the (1.55) as a recursive relation function of time imposing $\sum_k kn(k) \to 2t$ and $N_{kl}(t) \to tn_{kl}$:

$$\begin{aligned} n_{kl} &= \frac{4(l-1)}{k(k+1)(k+l)(k+l+1)(k+l+2)} \\ &+ \frac{12(l-1)}{k(k+l-1)(k+l)(k+l+1)(k+l+2)}. \end{aligned} \tag{1.56}$$

In general we can not factorize $n_{kl} = n_k n_l$ as in a random graph. Just in the case in which $1 \ll k \ll l$ we can do it getting:

$$n_{kl} \simeq k^{-2} l^{-2}, \tag{1.57}$$

that is different on respect of uncorrelated that give us $n_{kl} = k^{-3} l^{-3}$. This shows as the dynamical process that give us a scale-free topology correlations are introduced.

Clustering coefficient

No analytical result for the clustering coefficient has been found so far. It is easy to evaluate this coefficient numerically though. On respect to the random case for which $C_{rand} = \langle k \rangle / n$ the BA graph shows a different behavior: $C \sim n^{-0.7}$. In particular as show in Figure (1.2) in the BA the clustering coefficient is always higher.

1.5.4 Dorogovtsev-Mendes-Samukhin (DMS) model

In this model, (62), we have a preferential attachment criterion. In particular the probability that a new node i will be connected to a node j is

$$\pi(k_j, a) = \frac{a + k_j}{\sum_{l=1}^{i-1}(a + k_l)}. \tag{1.58}$$

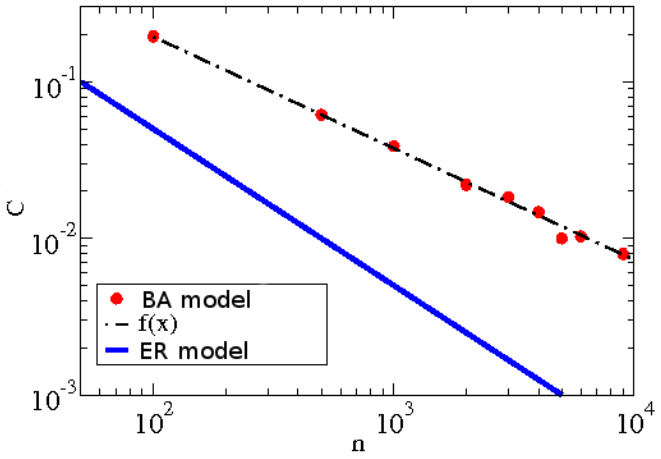

Figure 1.2: Average clustering coefficient as a function of the network size for BA model and ER model. $f(x) = n^{-0.7}$

It depends just on the positive constant a and on k_j. It is easy to prove how the degree distribution of the emergent graph is a scale free distribution with exponent $\gamma = 2 + a$. This model is just a generalization of the BA model that is found as a limit for $a = 1$.

2

Dynamical Processes on Complex Networks

Analogy pervades all our thinking, our everyday speech and our trivial conclusions as well as artistic ways of expression and the highest scientific achievements.

G. Pólya

Contents

In the next Chapters we will discuss about several different dynamical processes on complex networks. In particular we will be speaking about a class of these: Reaction-Diffusion processes. These are used to model a huge variety of phenomena in which local quantities obey physical reaction diffusion equations. Within the same general framework here introduced, we will describe the diffusion and localization of the *importance* of webpages and the spreading of infectious diseases. We will provide a general and abbreviated description of the theory that will be used in the remain parts of this work.

We invite the interested readers to classical textbooks for a more formal treatment of the subject (24; 66; 67; 68; 69; 70).

2.1 The master equation

Let us consider a network with N nodes. For each node i let us introduce a variable σ_i characterizing its dynamical state i.e. the evolution of particular attribute of that node. For each node we can enumerate all possible states $\sigma_i = 1, 2, \ldots, \kappa$, the knowledge of these state variables for all the vertices in the network defines the microstate of the whole system. For each time step t the system will be in some configuration generally described by set $\sigma(t) = (\sigma_1(t), \sigma_2(t), \ldots, \sigma_N(t))$. The dynamical evolution is defined by the dynamics of the configurations $\sigma(t)$ in the phase space. Let us considered σ^a and σ^b two different configurations of the system. The dynamical process is described by the transition $\sigma^a \to \sigma^b$. Large-scale systems have large number of variables and a stochastic nature. In general is not possible to follow the microscopic dynamics of it. For this reason it is important to focus on the probability $P(\sigma, t)$ of finding the system at time t in a given configuration σ and study its evolution through the *master equation*. In the continuous time approximation in a completely general fashion we can write:

$$\partial_t P(\sigma, t) = \sum_{\sigma'} \left[P(\sigma', t) W(\sigma' \to \sigma) - P(\sigma, t) W(\sigma \to \sigma') \right], \tag{2.1}$$

where the terms $W(\sigma' \to \sigma)$ represent the transition rates (hence they are unit $[\text{time}]^{-1}$) from one configuration to the other, the microscopic rules of the evolution are there encoded. In this general representation we have two terms only: one for the gain and one for the loss contributions for the probability distribution due to the transition from one state to the other. These rates are in general function of all configurations. It is quite common to consider cases in which the change of state of a node i is determined only by local interactions with its nearest neighbors. In this case we can write:

$$W(\sigma' \to \sigma) = \prod_i \omega(\sigma_i' \to \sigma_i | \sigma_j), \tag{2.2}$$

where the j are taken just in the set $\mathcal{V}(i)$ of neighbors of i. The role of the network is now clear. Its structure enters the dynamics since the transition rate for any node

depends on its neighborhood structure and then on network's topology.
Assuming to be able to solve the (2.1) is possible the calculation of the expectation values of all quantities of interest in the system. Considering a physical quantity A we have:

$$< A(t) >= \sum_{\sigma} A(\sigma)P(\sigma, t). \tag{2.3}$$

The average here is made considering all the states of the system, then it is called *phase space* average. In general these averages are function of time. A particular interesting case is the asymptotic limit:

$$\lim_{t\to\infty} P(\sigma, t) = P_{\infty}(\sigma). \tag{2.4}$$

This limit is just a mathematical concept. In real-world analysis, we are interested to the stationary state of the system that is reach when, after a typical transient time, the average over the stationary distribution is representative of the system.

2.1.1 Equilibrium and non-equilibrium systems

An isolated system maximizes its entropy and reaches a uniform stationary equilibrium distribution $P_{eq}(\sigma)$ with the same probability of being in any of the fixed energy accessible configurations. Isolated systems are characterized by the fact that average over the time evolution of any quantity of interest is the same as the average over the stationary equilibrium distribution. These are the entropy maximization axiom and the ergodic hypothesis. Real system are never isolated. They are always coupled with the external environment, that we can consider as a heat bath that fix the equilibrium temperature of the system. In this case the stationary distribution is no longer uniform but characterized by the Boltzmann-Gibss distribution:

$$P_{eq}(\sigma) = \frac{1}{Z} e^{-H(\sigma/k_B T)}. \tag{2.5}$$

Where T is the temperature, k_B is the Boltzman factor, $H(\sigma)$ is the Hamiltonian of the system (that gives the energy associated to each configuration of the system) and Z the partition function:

$$Z = \sum_{\sigma} e^{-H(\sigma)/k_B T}, \tag{2.6}$$

that gives the correct normalization factor. The (2.5), in case of equilibrium physical systems, may be obtained just by knowing the system Hamiltonian.

The equilibrium state is characterized by the detailed balance condition on the master equation: the net probability current between pairs of configurations is zero when $P = P_{eq}$

$$P_{eq}(\sigma)W(\sigma \to \sigma') = P_{eq}(\sigma')W(\sigma' \to \sigma). \tag{2.7}$$

This is a strong condition that implies that each pairs of terms in the master equation has a null contribution. This is not true in the case of systems out of equilibrium, where currents between microstates do not balance. This is due to the fact that systems could be not isolated and are subject to external currents or driving forces such as addition of energy and particles or the presence of dissipation and therefore out of equilibrium. Many of these systems are characterized by the presence of absorbing state, configurations that can only be reached but not left. In this case we always have a non-zero probability current for some some configurations so that the temporal evolution cannot be described by an equilibrium distribution.

2.1.2 Approximate solutions of the master equation

As we already said, even for simple dynamical processes a complete solution of the master equation can be rarely derived. There are standard approximation schemes that we will present in this section that are used to find proxy of the general solution.
The first thing to do is to consider appropriate projections focusing our interest on specific quantities. For example we can consider the average number of nodes in the state x at time t:

$$< N_x(t) > \equiv N_x(t) = \sum_{\sigma,i} \delta_{\sigma_i,x} P(\sigma, t). \tag{2.8}$$

This is general not enough for a good representation of the system. A further approximation scheme is the homogeneous assumption or mean-field theory. In this approach the system is considered homogeneous and the correlations between microstate variables are neglected. The interactions felt by any element in the system are the same, and they can be thought as an average interaction due to the full system. In other words the probability for a given i to be in the state $\sigma_i = x$ is a p_x independent of i. So, neglecting the correlations we can write the probability of any configuration as a factorization of single node probabilities:

$$P(\sigma) = \prod_i p_{\sigma_i}. \tag{2.9}$$

Using these consideration we can write

$$\partial_t N_x(t) = F_x(N_1, N_2, \ldots, N_\kappa). \tag{2.10}$$

The explicit form of the function F_x is related on the specific interactions among nodes, transition rates and number of allowed states.

2.1.3 $A + B \to 2B$ process, mean-field approach

Let us consider as example a fundamental process in which each node can be only in two states $\sigma_i = \{A, B\}$. Let us fix the dynamical rule of the process:

$$A + B \to 2B, \tag{2.11}$$

in words A can be converted in B just interacting with another B and the process is irreversible and occurs with rate β. The transition rates are then:

$$\omega(A \to A|\sigma_j = A) = \omega(B \to B|\sigma_j = A) = \omega(B \to B|\sigma_j = B) = 1, \tag{2.12}$$

and

$$\omega(A \to B|\sigma_j = B) = \beta. \tag{2.13}$$

To define deterministic equations we use the quantities:

$$N_A(t) = \sum_{\sigma,i} \delta_{\sigma_i,A} P(\sigma, t), \quad N_B(t) = \sum_{\sigma,i} \delta_{\sigma_i,B} P(\sigma, t). \tag{2.14}$$

We can now write for B:

$$\begin{aligned} \partial_t N_B(t) &= \sum_{\sigma,i} \delta_{\sigma_i,B} \partial_t P(\sigma, t) \qquad (2.15) \\ &= \sum_{i,\sigma',\sigma} \left[\delta_{\sigma_i,B} \prod_k \omega(\sigma'_k \to \sigma_k|\sigma'_j) P(\sigma', t) - \delta_{\sigma_i,B} \prod_k \omega(\sigma_k \to \sigma'_k|\sigma'_j) P(\sigma, t) \right], \end{aligned}$$

that after a bit of algebra reads:

$$\partial_t N_B(t) = \sum_{i,\sigma'} \left[\omega(\sigma'_i \to \sigma_i = B|\sigma'_j) P(\sigma', t) \right] - N_B(t). \tag{2.16}$$

It is time to introduce the mean-field approximation stating that the probability for each node to be in the state A or B is $p_A = N_A/N$ and $P_B = N_B/N$. Even more neglecting correlation we can write:

$$\begin{aligned}\sum_{\sigma'} \omega(\sigma'_i \to \sigma_i = B|\sigma'_j) P(\sigma', t) &= \\ &= \sum_{\sigma'_j} [\omega(\sigma'_i = A \to \sigma_i = B|\sigma'_j) P(\sigma', t) p_A \prod_{j \in \mathcal{V}(i)} p_{\sigma'_j} + \\ &+ \omega(\sigma'_i = B \to \sigma_i = B|\sigma'_j) P(\sigma', t) p_B \prod_{j \in \mathcal{V}(i)} p_{\sigma'_j}],\end{aligned} \tag{2.17}$$

where the sum is run just on the nearest neighbors is of j. We can simplify more considering that $\omega(\sigma'_i = B \to \sigma_i = B|\sigma'_j) = 1$ independently of the configuration of j and that $\omega(\sigma'_i = A \to \sigma_i = B|\sigma'_j) = \beta$ if at least one of the k neighbors of i is in the state B. This happen with probability $1 - (1 - p_B)^k$. Using this and assuming that the k is the same for all the nodes we get:

$$\partial_t N_B(t) = \sum_i \Big[\beta p_A (1 - (1 - p_B)^k) + p_B\Big] - N_B(t), \tag{2.18}$$

or

$$\partial_t N_B(t) = \beta N_A \left[1 - \left(1 - \frac{N_B}{N}\right)^k\right], \tag{2.19}$$

that in the limit $N_B/N \ll 1$ yields the dynamical equation:

$$\partial_t N_B(t) = \beta k \frac{N_A N_B}{N}. \tag{2.20}$$

The equation for A is given by the conservation rule $N_A = N - N_B$. The expression (2.20) is the mean-field solution for the basic reaction process $A + B \to 2B$ that is part of a wide range of epidemic spreading phenomena that we will discuss in detail in the following sections.

2.1.4 Diffusion processes and random walk

One of the most important dynamical processes are the diffusion processes. The application that we will discuss in details in this work are related to exploration and information retrieval form the vertices of a network, and modeling infectious diseases.

The simplest strategy to explore a network is to choose randomly a starting node, follow randomly one of its connections and iterate this process until a satisfactory knowledge of

the network is found: this is a random walk. Let us consider W walkers in a undirected network with N nodes and degree distribution $P(k)$. We can define for each node an occupation number W_i that gives us the number of walker present at each time step. Of course

$$W = \sum_i W_i. \tag{2.21}$$

The diffusion of these walkers is characterized by the topology of the network but even by the transition rates that define how each walker diffuses along the edges. The simplest hypothesis that we can do is to assume that the movement at the time step t does not depends on all the history of each walker but just on the state at the time step $t-1$: Markovian processes. The transition rate then be can be written as:

$$d_{ij} = \frac{r}{k_i}, \tag{2.22}$$

where k_i is the degree of i. Each link connected to i has the same probability to be selected and we will have a total rate of escape:

$$r = \sum_{j \to i} d_{ij}. \tag{2.23}$$

As we will see in other sections for a statistical characterization of networks it is convenient to group the nodes in degree classes. We are assuming then a statistical equivalence of nodes with the same degree. This is a classic approach in many dynamical process such as epidemics and opinion dynamics. In this case we can consider the average number of walkers in nodes within the degree k:

$$W_k = \frac{1}{N_k} \sum_{i|k_i=k} W_i, \tag{2.24}$$

where N_k is the number of nodes with degree k. Using again mean-field dynamical equation for the variation in time of $W_k(t)$ we can write:

$$\partial_t W_k(t) = -rW_k(t) + k \sum_{k'} P(k'|k) \frac{r}{k'} W_{k'}(t). \tag{2.25}$$

The first term on the right side considers the escape rate r. The second term takes into account the walkers diffusing in from all neighbors. This term is proportional to the

number of nodes in the neighbors k times the average number of walkers coming from each neighbor. In case of uncorrelated networks we can write:

$$P(k'|k) = k' \frac{P(k')}{< k >}, \tag{2.26}$$

then we have:

$$\partial_t W_k(t) = -rW_k(t) + \frac{k}{< k >} \sum_{k'} P(k') r W_{k'}(t). \tag{2.27}$$

The stationary condition $\partial_t W_k(t) = 0$ does not depend upon the diffusion rate r that fixes the time scale at which the equilibrium is reached and has the solution

$$W_k = \frac{k}{< k >} \frac{W}{N}, \tag{2.28}$$

where $W/N = \sum_{k'} P(k') W_{k'}(t)$ is the average number of walkers per node that is constant. We can now define the probability to find a single diffusing walker in a node of degree k, $p_k = W_k/W$ obtaining:

$$p_k = \frac{k}{< k >} \frac{1}{N}. \tag{2.29}$$

The stationary visiting probability of a random walker in an uncorrelated network with arbitrary degree distribution is proportional to the degree.

Another classical measure for a diffusion process is the return probability $p_0(t)$. This gives the probability that a walker returns to its starting point after t steps. As shown in details in the Ref. (24) this quantity is related to the spectral density ot the modified Laplacian operator associated with this process:

$$L_{ij} = \delta_{ij} - \frac{x_{ij}}{k_j}. \tag{2.30}$$

In particular we have:

$$p_0(t) = \int_0^\infty d\lambda e^{-\lambda t} \rho(\lambda). \tag{2.31}$$

Interestingly the behavior in the long time limit is related to the behavior of the spectral density for $\lambda \to 0$. For D-dimensional lattice we have (71):

$$p_0(t) \sim t^{-D/2}. \tag{2.32}$$

For a ER graph (72):

$$p_0(t) \sim e^{at^{1/3}}, \tag{2.33}$$

where a is a constant related on the specific network. For a WS graph (73):

$$p_0(t) - p_0(\infty) \sim \begin{cases} t^{-D/2} & \text{if } t \ll t_1 \\ e^{-(t/t_1)^{1/3}} & \text{if } t_1 \ll t \end{cases} \tag{2.34}$$

where $p_0(\infty) = 1/N$ and the $t_1 \sim 1/p^2$. In the first regime, the diffusion is the same as on D-dimensional lattice, but after time order t_1 the walkers will start to feel the effect on the shortcuts typical of the graph. This regime is characterized by the behavior observed in a ER graph.
A recent work based on uncorrelated random scale-free networks with minimum degree m equal to one or two (74) shows that:

$$p_0(t) \sim t^{\eta} e^{-bt^{1/3}}, \tag{2.35}$$

that is different from the WS case. η and b are constants related to the specific network. Another important quantity is the average time to visit or to return to a node. This is inversely proportional to its degree (75). This is another case in which the importance of hubs and scale-free topology have an huge impact on the dynamics. It is possible a generalization of all the quantity here introduced for directed networks and weighted networks. For the details we invite the reader to the Chapters 4 and 7.

2.2 Epidemic spreading on complex networks

Using the general framework of dynamical processes on complex networks that we have introduced we will study another application of those concepts: basic epidemic models. The simplest class assumes that the population is divided into different compartments depending on the stage of the disease (76; 77) such as susceptible S, those who can contract the infection, infectious I, those who have already contracted the infection and recovered R, those who have recovered from the disease. Additional stages of the disease can be introduced depending on the type of the disease. Examples of these extensions will be shown in details in the Chapters 5 and 7. Let us consider a population of N individuals and let us define the number of individuals in the class $[m]$ at the time t as $X^{[m]}(t)$. Assuming a conservation of the number of people

$$N = \sum_m X^{[m]}(t). \tag{2.36}$$

The transitions between different compartments depends on the specific disease that we are modeling. The rate at which susceptible contract the infection (force of infection) has generally two possible form (78):

- frequency dependent or mass action transmission (equivalent to a homogeneous or mean-field approximation)
- density dependent or pseudo mass action.

The first kind reflects the situation where the number of contacts is independent of the population size but just on the fraction of infectious individuals in the population. The second kind instead assumes that as the population size increase so does the contact rate. These kind of processes are anyway binary interactions among individuals, so processes that we already introduced:

$$A + B \to 2B. \tag{2.37}$$

Considering mass action transmission the variation of $X^{[m]}$ due to this process is given by $\sum_{h,g} \nu^m_{h,g} a_{h,g} X^{[h]} X^{[g]} N^{-1}$, where $\nu^m_{g,h} = [-1, 0, 1]$ and $a_{h,g}$ is the transition rate of the process.
Typically another kind of process is considered: a spontaneous transition of one individual from one compartment $[m]$ to another one $[h]$. Processes of this kind can be used to model spontaneous recovery of infected individuals and usually described as:

$$B \to A. \tag{2.38}$$

The variation in the number of individuals $X^{[m]}$ is simply given by $\sum_h \nu^m_h a_h X^{[h]}$ where $\nu^m_h = [-1, 0, 1]$ and a_h is the transition rate. We can now write the general deterministic reaction rate equations for the quantity $X^{[m]}$ summing the two contribution presented:

$$\partial_t X^{[m]} = \sum_{h,g} \nu^m_{h,g} a_{h,g} X^{[h]} X^{[g]} N^{-1} + \sum_h \nu^m_h a_h X^{[h]}. \tag{2.39}$$

Within this framework we can easily derive the dynamical equations for three basic models: SI, SIS and SIR.

2.2.1 SI model in homogeneous networks

In this model every node can only exist in two discrete states, susceptible or infected. The probability that a susceptible acquires the infection form any given neighbor in a time interval dt is βdt where β is the pathogen spreading rate. In this model individuals that enter in the I state remain permanently infectious. The $I(t)$ or $i(t) = I(t)/N$ can just increase over time. Every infected node attempts to infect a connected susceptible vertex with probability βdt. The probability of getting infected having n infected neighbors is $1 - (1 - \beta dt)^n$. Considering on average k contacts for each individuals and assuming $\beta dt \ll 1$ we can write:

$$1 - (1 - \beta dt)^{ki} \simeq \beta k i dt. \tag{2.40}$$

The evolution of the SI using the general (2.39) reads as:

$$d_t i(t) = \beta < k > i(t)[1 - i(t)], \tag{2.41}$$

that is exactly what we got in the (2.20). Of course $1 - i(t) = s(t)$.

2.2.2 SIS model in homogeneous networks

In this model individual exist in two class only as in the previous model. The disease transmission is described as in the SI model but infected individuals may recover and come back in the susceptible class again due to a spontaneous transition with probability μdt, where μ is the recovery rate. Using the general (2.39) for this processes we have:

$$d_t i(t) = -\mu i(t) + \beta < k > i(t)[1 - i(t)]. \tag{2.42}$$

The cycle susceptible $\rightarrow$ infected $\rightarrow$ susceptible can lead to an endemic state with a stationary and constant number of infected individuals.

2.2.3 SIR model in homogeneous networks

In this model the infected individuals recover with rate μ and enter a new compartments R of removed individuals (79). Using the general equation (2.39) for our three compartments we get:

$$d_t s(t) = -\beta < k > i(t)[1 - r(t) - i(t)], \tag{2.43}$$

for S,

$$d_t i(t) = -\mu i(t) + \beta < k > i(t)[1 - r(t) - i(t)], \tag{2.44}$$

for I,

$$d_t r(t) = \mu i(t), \tag{2.45}$$

for R. In this model as in the SIS model we have a time scale μ^{-1} governing the self-recovery of individuals. There are two competing processes: the infection and the recovery. If $\mu^{-1} \ll \beta^{-1}$ the recovery of individuals is much faster and the system decay into a healthy state. Instead if $\mu^{-1} \gg \beta^{-1}$ the spreading time scale is much smaller than the recovery time scale. The recovery will occur in a later stage on respect to the early dynamics of the epidemic outbreak.

2.3 Epidemic threshold

All the models we defined so far can be easily solved at the early stage of the epidemics when we can assume that the number of infected individuals is very small fraction of the whole population. We can solve the differential equations in the limit $i(t) << 1$ with a linear approximation neglecting all the terms order $\mathcal{O}(i^2)$. The equation of infected for the SI model reads as:

$$d_t i(t) \sim \beta < k > i(t), \tag{2.46}$$

with solution

$$i(t) \sim i_0 e^{\beta < k > t}, \tag{2.47}$$

where i_0 is the initial density of infected individuals. It is clear, since in the exponential all the factors are positive, the epidemic always propagates in the population infecting all the individuals. For this basic model an exact complete solution is easily obtain:

$$i(t) = \frac{1}{1 + i_0(e^{\frac{t}{\tau}} - 1)} i_0 e^{\frac{t}{\tau}}, \tag{2.48}$$

that recovers the (2.47) for $t \ll \tau$ and gives $i \to 1$ for $t \gg \tau$ and where $\tau = (\beta < k >)^{-1}$. Using the same linear approximation we can find an expression for the SIS and SIR model:

$$d_t i(t) \sim \mu i(t) + \beta < k > i(t). \tag{2.49}$$

The solution of this differential equation is:

$$i(t) \sim i_0 e^{\frac{t}{\tau}}, \tag{2.50}$$

where we set:

$$\tau^{-1} = \beta < k > -\mu. \tag{2.51}$$

In this case the argument of the exponential is not always positive. If

$$\beta < \frac{\mu}{< k >}, \tag{2.52}$$

we get a negative term. In this case we have an exponential decay of the fraction of infected individuals. The epidemic outbreak will not affect a finite portion of the population and will dies out in a finite time. We have then an *epidemic threshold*:

$$\tau^{-1} = \mu(R_0 - 1) > 0, \tag{2.53}$$

where $R_0 = \beta < k > /\mu$ is the basic reproductive rate in the SIS and SIR model. The spreading will occur provided $R_0 > 1$. To be more precise the stochastic fluctuations may lead to the extinction of the epidemics even if the system is above the threshold. As shown in Ref. (80) the extinction probability of a epidemic starting with i_0 infected individuals is:

$$P^{ext} = \frac{1}{R_0^{i_0}}. \tag{2.54}$$

The concept of epidemic threshold is very general and a key property of all epidemics models. The expression may be different for different models but still present.
We can defined in general three different stages of the epidemic evolution: pre-outbreak, exponential growth and final. The first stage is dominated by stochastic effects due to the small number of infected individuals. This stage cannot be described by the deterministic continuous equations that we derived in the homogeneous approximation. A full stochastic analysis would be required. The epidemics may or not disappear just due to stochastic effects. If the epidemic survives to the first stage and the fraction of infected individuals is enough to make stochastic effect negligible but small enough on respect to the whole population we enter in the exponential stage. The final stage is model dependent. In general we can say that the decrease of susceptible slow down the growth of infected individuals and the exponential stage is not possible any longer. In the SI the number of infected will keep increasing at lower rate until the total population

is infected. In the SIS model the fraction of infected will enter in a stationary state fixed by the dynamical balance between spreading and recovery rate. In the SIR model the number of infected will decrease and the disease will die out due to the increase of the recovered individuals.

2.4 Epidemics in heterogeneous networks

As we said in the Chapter 1 many real social and technological networks of epidemiological relevance (mobility networks, the web of sexual contacts and internet) are far to be homogeneous. The hypothesis that each individual in the system has the same number of connections $k \simeq < k >$ (that we used in the previous sections) is not a good approximation. The fluctuations play a main role in determining the epidemic properties and the spreading may be favored in heterogeneous networks (21; 32; 76; 81).
Even in this case we will consider a degree block approximation: all nodes with the same degree are statistically equivalent. The quantity we will study are:

$$i_k = \frac{I_k}{N_k}, \quad s_k \frac{S_k}{N_k}. \tag{2.55}$$

The global averages are given by:

$$i = \sum_k P(k) i_k \quad , s = \sum_k P(k) s_k. \tag{2.56}$$

2.4.1 The SI model

In this case we know that the system will be totally infected independently of the spreading rate, but it is very interesting to see the effect of topological fluctuations on the spreading velocity. Considering the class of degree k and defining $\theta_k(t)$ the density of infected neighbors of vertices of degree k the evolution equations read:

$$d_t i_k(t) = \beta[1 - i_k(t)] k \theta_k(t). \tag{2.57}$$

In the homogeneous assumption the last term was equal to the density of infected nodes. In a heterogeneous network it is in general a very complicated term that takes into into account the different degree classes and their connections. The simplest case we can analyzed is a network with no degree correlations:

$$\theta_k(t) = \theta(t) = \frac{\sum_{k'} (k' - 1) P(k') i_{k'}(t)}{< k >}. \tag{2.58}$$

Using this in the (2.57) we have:

$$d_t i_k(t) = \beta k \theta(t), \tag{2.59}$$

multiplying both sides of this expression for $\sum_k (k-1)P(k)$ and summing over k we get:

$$d_t \theta(t) = \beta \theta(t) \left(\frac{< k^2 >}{< k >} - 1 \right). \tag{2.60}$$

We can solve this two coupled equations fixing $i_k(t=0) = i_0$ getting:

$$i_k(t) = i_0 \left[1 + \frac{k(< k > -1)}{< k^2 > - < k >} (e^{t/\tau} - 1) \right], \tag{2.61}$$

with

$$\tau = \frac{< k >}{\beta(< k^2 > - < k >)}. \tag{2.62}$$

It is clear that the fraction of infected individuals increases exponentially. This processes is fast for high degree nodes. The growth time scale is measured by the heterogeneity ratio $< k^2 > / < k >$. For scale free networks with exponent $2 < \alpha \leq 3$ in the limit $N \to \infty$ we have an unbounded second moment, then in uncorrelated scale-free networks we would have a virtually instantaneous rise of the epidemic size. The reason for that is quite intuitive. Once the disease has reached the hubs it can spread rapidly among the network. Multiplying both sides for $P(k)$ and summing over all k we get:

$$i(t) = i_0 \left[1 + \frac{< k >^2 - < k >}{< k^2 > - < k >} (e^{t/\tau} - 1) \right]. \tag{2.63}$$

In the case of presence of non trivial correlations we have the general expression (82):

$$\theta_k = \sum_{k'} i_{k'} \frac{k'-1}{k'} P(k'|k). \tag{2.64}$$

Neglecting terms of order $\mathcal{O}(i^2)$ we get for $i_k(t)$:

$$d_t i_k(t) = \sum_{k'} \beta k \frac{k'-1}{k'} P(k'|k) i_{k'}(t) \equiv \sum_{k'} C_{k,k'} i_{k'}(t), \tag{2.65}$$

a linear system of differential equation given by the matrix $C = C_{k,k'}$. The solution will be a linear combination of exponential functions of the forms $e^{\Lambda_i t}$ where Λ_i are eigenvalues of the matrix C. We can approximate

$$i(t) \sim e^{\Lambda_m t}, \tag{2.66}$$

using the largest eigenvalue Λ_m. Using the Frobenius theorem the largest eigenvalue for correlated networks is bounded form below (83):

$$\Lambda_m^2 \geq \min_k \sum_{k',l} (k'-1)(l-1)P(l|k)P(k'|l), \tag{2.67}$$

that we can rewrite as:

$$\Lambda_m^2 \geq \min_k \sum_{l} (l-1)P(l|k)(k_{nn}(l)-1). \tag{2.68}$$

For scale-free networks with exponent $2 < \alpha \leq 3$ in the limit of infinite size k_{nn} is divergent which implies that the largest eigenvalues is unbounded too.

2.4.2 The SIS and SIR model

A generalization for these two model is quite easy:

$$d_t i_k(t) = \beta k s_k(t)\theta_k(t) - \mu i_k(t), \tag{2.69}$$

where $s_k(t) = 1 - i_k(t)$ for the SIS model and $s_k(t) = 1 - r_k(t) - i_k(t)$ for the SIR model. Again considering the linear approximation and uncorrelated networks we get the time scale τ:

$$\tau = \frac{< k >}{\beta < k^2 > -(\mu + \beta) < k >}. \tag{2.70}$$

We have here a threshold. In order to ensure an epidemic outbreak the condition $\tau > 0$ must be satisfy:

$$\frac{\beta}{\mu} \geq \frac{< k >}{< k^2 > - < k >}. \tag{2.71}$$

For scale-free networks with exponent $2 < \alpha \leq 3$ in limit of infinite size the second moment diverges, so we have a null epidemic threshold. In real cases the threshold is not zero, but really small. This is an important result that confirm how heterogeneous networks behave in a completely different way from homogeneous networks. Scale-free networks are then an ideal topology for the spreading of infectious diseases. Fortunately the prevalence for small spreading rates is very small and as we will see the heterogeneity could be an advantage in vaccination strategies of great effectiveness.

2.4.3 $t \to \infty$ limit

We studied the early time regime so far. The other limit $t \to \infty$ is also very interesting. For the SI model this limit is trivial, we know in fact how $i(t \to \infty) = 1$. For the other two models the situation is very different.

SIS model

Let us consider the case of a generalized random graph with no degree correlations. We already said that in this case the function $\theta_k = \theta$ is independent from k. In the limit $d_t i_k(t) = 0$ from the (2.69) we get:

$$i_k = \frac{\beta k \theta}{\mu + k\beta\theta}. \tag{2.72}$$

Using this into (2.58) we get a self consisted equation:

$$\theta = \frac{1}{< k >} \sum_k (k-1) P(k) \frac{\beta k \theta}{\mu + k\beta\theta}. \tag{2.73}$$

We can explicitly calculate the epidemic threshold from this equation as shown in Ref. (81) just considering that the condition is given by the value of β and μ for which it is possible to obtain a non-zero solution θ^*. Using geometrical considerations (24) we get:

$$d_\theta \left(\frac{1}{< k >} \sum_k (k-1) P(k) \frac{\beta k \theta}{\mu + k\beta\theta} \right) |_{\theta=0} = \frac{\beta < k^2 >}{\mu < k >} \geq 1, \tag{2.74}$$

then an epidemic threshold condition that recovers the results obtained form the linear approximation:

$$\frac{\beta}{\mu} = \frac{< k >}{< k^2 >}. \tag{2.75}$$

Let us focus our attention on random uncorrelated scale-free networks defined by:

$$P(k) = (\alpha - 1) m^{\alpha - 1} k^{-\alpha}, \tag{2.76}$$

where m is the minimum degree of any vertex. It is trivial to evaluate the moments of this distribution and using (2.75) we get:

$$\frac{\beta}{\mu} = \begin{cases} \frac{\alpha - 3}{m(\alpha - 2)} & \text{if } \alpha > 3 \\ 0 & \text{if } \alpha \leq 3 \end{cases} \tag{2.77}$$

To get the density of infected individuals in the stationary states we have to solve explicitly the self consistent equation for θ in the limit of β/μ approaching the epidemic threshold (81). The results are different for different value of the exponent α. For $2 < \alpha < 3$ we get:

$$i_\infty \sim \left(\frac{\beta}{\mu} \right)^{\frac{1}{3-\alpha}}, \tag{2.78}$$

it is worth to notice that exponent is larger than 1, this implies that for small β/μ the number of infected individuals is growing very slowly. There is a wide region of spreading rates in which $i_\infty \ll 1$. There is no an epidemic threshold as was aspected. For $\alpha = 3$ we have:

$$i_\infty \sim e^{-\frac{\mu}{m\beta}}, \tag{2.79}$$

we have, even in this case, an absence of epidemic threshold and for a wide range of spreading rates in which $i_\infty \ll 1$. For $3 < \alpha < 4$ we have:

$$i_\infty \sim \left(\frac{\beta}{\mu} - \frac{\alpha - 3}{m(\alpha - 2)}\right)^{\frac{1}{\alpha-3}}, \tag{2.80}$$

a power law behavior is observed associated with a non-zero threshold. This threshold is approached without any sign of the singular behavior usually associated to a critical point. For $\alpha > 4$ we have:

$$i_\infty \sim \frac{\beta}{\mu} - \frac{\alpha - 3}{m(\alpha - 2)}, \tag{2.81}$$

that is the usual epidemic threshold we found for homogeneous networks.

SIR model

As we said in the SIR model $i_\infty = 0$. The epidemics dies due to the depletion of the susceptible individuals that after recovering move into the removed compartment. Another interesting quantity is provided by the total number of individuals affected by the infection: $r_\infty = \lim_{t\to\infty} r(t)$. Let consider the system of differential equation for the SIR model. Let us focus in the early time in the case in which $i_k(0) = i_0 \simeq 0$ and $s_k(0) \simeq 1$. We can integrate directly the equation getting:

$$s_k(t) = e^{-\beta k \phi(t)}, \quad r_k(t) = \mu \int_0^t d\tau i_k(\tau), \tag{2.82}$$

where we used

$$\phi(t) = \int_0^t d\tau \theta(\tau) = \frac{1}{< k > \mu} \sum_k (k-1) P(k) r_k(t). \tag{2.83}$$

Taking the derivate of both members of this we get:

$$\begin{aligned} d_t \phi(t) &= \frac{1}{< k > \mu} \sum_k (k-1) P(k) [1 - r_k(t) - s_k(t)] \\ &= 1 - \frac{1}{< k >} - \mu\phi(t) - \frac{1}{< k >} \sum_k (k-1) P(k) e^{-\beta k \phi(t)}. \end{aligned} \tag{2.84}$$

This equation can not be generally solved, but we can get information on the infinite time limit. Due to the conservation of the number of individuals in the system:

$$1 = s_k(\infty) + r_k(\infty), \tag{2.85}$$

so we have:

$$r_\infty = \sum_k P(k)(1 - e^{\beta k \phi_\infty}). \tag{2.86}$$

In the infinite time limit $d_t\phi = 0$ then:

$$\mu\phi_\infty = 1 - \frac{1}{< k >} - \frac{1}{< k >}\sum_k (k-1)P(k)e^{-\beta k\phi_\infty}. \tag{2.87}$$

The value $\phi_\infty = 0$ is a trivial solution. The non trivial solution is related to $r_\infty > 0$ and exist only if:

$$d_{\phi_\infty}\left(1 - \frac{1}{< k >} - \frac{1}{< k >}\sum_k (k-1)P(k)e^{-\beta k\phi_\infty}\right)|_{\phi_\infty=0} \geq \mu, \tag{2.88}$$

that is equivalent to:

$$\frac{\beta}{< k >}\sum_k k(k-1)P(k) \geq \mu. \tag{2.89}$$

This defines the well known epidemic threshold condition:

$$\frac{\beta}{\mu} > \frac{< k >}{< k^2 > - < k >}. \tag{2.90}$$

Expressions of r_∞ are not easy to be found, in some case such as heavy-tailed networks of finite size and small values of β/μ we have $r_\infty \sim (\beta/\mu)^{1/(3-\alpha)}$ for $2 < \alpha < 3$ and $r_\infty \sim e^{-\mu/m\beta}$ for $\alpha = 3$ (84; 85).

2.4.4 Immunization

From the previous sections we saw how many real networks of epidemic relevance are heterogeneous and that in these networks the epidemic threshold are extremely small. This is a worrying scenario. However it is possible to take advantage of the heterogeneity developing new defensive strategies extremely effective. In this section we will give an brief overview of the principle method and results.

A possible way of immunize people is the uniform distribution. These method is completely inefficient in heterogeneous networks because it gives that same importance to

vertices with a very small degree and to vertices with a large connectivity. The introduction of a fraction g of immune individuals chosen at random is equivalent to a rescaling of the effective spreading rate:

$$\beta \to \beta(1-g), \tag{2.91}$$

the rate at which new infected individuals appear is depressed by a factor proportional to the probability that they are not immunized. For uncorrelated networks we have:

$$\frac{\beta}{\mu}(1-g_c) = \frac{< k >}{< k^2 >}, \tag{2.92}$$

where g_c is the immunization threshold. In the case of heavy-tailed networks with a diverging second moment in the thermodynamic limit only a complete immunization of the networks ensures an infection-free stationary state. This strategy is not effective.
Scale-free network are strongly affected by targeted damage. If a few of the most connected nodes are removed the network suffers a huge reduction of its ability to carry informations (24). This can be extremely helpful in case of the spreading of infectious diseases. A targeted immunizations in which we progressively make immune the hubs will be very effective since the principal actors in the spreading will be blocked. Let us consider, following Ref. (86), that a fraction g of the individuals with the highest degree have been immunized. This is an introduction of an upper cut-off $k_c(g)$ function of g such that all nodes with degree $k > k_c(g)$ are immune. Immunization of nodes means that the infection can not propagate along all the edges emanating from these nodes. The elimination of nodes and links for the spreading purpose yields a new topology with moments $< k >_g$ and $< k^2 >_g$ can be evaluated as a function of the density of immunized individuals (86). The protection of the network will be achieved when the effective network on which the epidemic spreads satisfies:

$$\frac{< k >_{g_c}}{< k^2 >_{g_c}} \geq \frac{\beta}{\mu}, \tag{2.93}$$

that gives the immunization threshold

$$\frac{< k >_{g_c}}{< k^2 >_{g_c}} = \frac{\beta}{\mu}. \tag{2.94}$$

For uncorrelated scale-free network with exponent $\alpha = 3$ it is possible to perform the explicit calculation (86) getting:

$$g_c \sim e^{-2\frac{\mu}{m\beta}}, \tag{2.95}$$

where m is the minimum degree of the network. This result shows clearly that targeted immunization is extremely convenient, with a critical immunization threshold that is exponentially small in a wide range of spreading rates.

2.5 Single population and homogeneous mixing

As we discussed in the previous sections the heterogeneities plays a crucial role in the spread of epidemic diseases. For some type of diseases, such for example the flu, the process of contagion does not require a personal contact as in the case of HIV or other sexually transmitted diseases. It is possible to get the flu just staying in the same classroom, bus or office; during all our normal activities. We can imagine than within a community each susceptible individual can meet an infected one with a probability proportional to the ratio of infected individuals. This is the homogeneous mixing approximation that can be used in absence of detailed data of the connectivity patterns to model diseases in which the contagion process does not involve personal face to face interactions.
In this section we discuss this approximation that will used extensively in Chapter 5, 6 and 7.
Let us imagine to have a population of N individuals. Let us consider a disease that can be modeled with the homogeneous mixing assumption.

2.5.1 SI model

In this model we have just two possible compartments: S and I. At any time t each susceptible individual with probability $\frac{I}{N}$ can get in touch with an infected one. The probably the a susceptible acquires the infection from each interaction in a time interval dt is βdt, where β defines the pathogen spreading rate of the disease. Of course any different disease is characterized by a different β. At every time step the average number of interaction between S and I is $S\frac{I}{N}$. then the total number S that becomes infected is $\beta dt S\frac{I}{N}$. Then the number of S individual after dt is:

$$S(t+dt) = S(t) - \beta S\frac{I}{N}dt, \tag{2.96}$$

and for I

$$I(t+dt) = I(t) + \beta I\frac{S}{N}dt. \tag{2.97}$$

We can make the limit for $dt \to 0$[1] and get a system of partial non linear differential equation:

$$\partial_t S = -\beta S \frac{I}{N}, \tag{2.98}$$

$$\partial_t I = \beta I \frac{S}{N}. \tag{2.99}$$

We can solve the equation using the initial conditions:

$$\begin{aligned} S(t_0) &= S_0, \\ I(t_0) &= I_0, \\ N &= I_0 + S_0 = I(t) + S(t). \end{aligned} \tag{2.100}$$

Using the last condition in the equation for I considering $N = 1$ or S and I as the density ratio/density of individuals we get:

$$\partial_t I = \beta(1 - I)I, \tag{2.101}$$

then

$$\int_{I_0}^{I} \frac{dI'}{I'(1 - I')} = \beta \int_{t_0}^{t} dt, \tag{2.102}$$

that we can intrate easily getting:

$$I(t) = \frac{1}{1 + \frac{a}{b} e^{-\beta t}}, \tag{2.103}$$

and of course:

$$S(t) = 1 - I(t), \tag{2.104}$$

where $a = I_0^{-1} - 1$and $b = e^{-\beta t_0}$. It is clear from the equation (2.103) that for $t \to \infty$:

$$S_\infty = 0; \quad I_\infty = 1. \tag{2.105}$$

Basically just with one seed (one infect individual) all the population become infected. That because the unique possible interaction is:

$$S + I \to 2I, \tag{2.106}$$

[1]The limit for infinitesimally small time of interaction is just a mathematical modelization of the process. Of course the is a finite time scale for the interaction

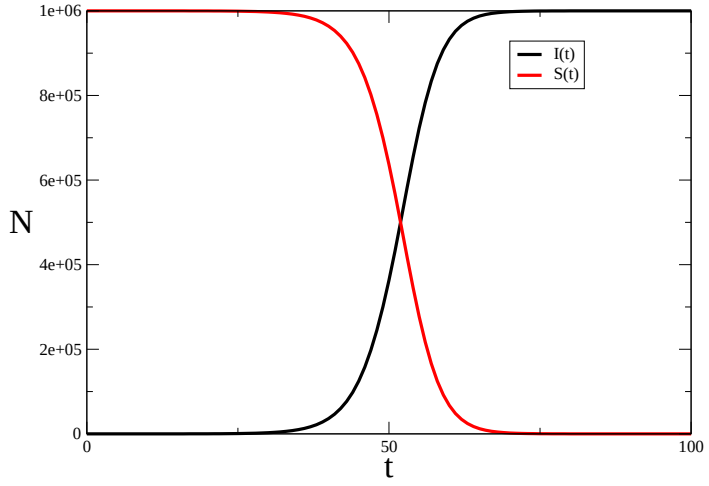

Figure 2.1: SI model's curves (median) for $\mu = 0.1$ and $R_0 = 3$ and 1000 realizations.

and it will be active as long as the number of S will reach 0. In Figure (2.1) we show a typical example of SI's curves I and S. We used a population of $N = 10^6$ with 1 seed ($I_0 = 1$). It is clear that the two profiles are totally reciprocal, since the rate of decrease of one (S) is equal to the rate of increase of the other I. The profile presented are the median of 1000 realizations. Since we modeled the simulation through a stochastic dynamic we need a sufficient number of simulations to reduce the effect of fluctuations.

2.5.2 SIS model

This model is a more realist one, used to model the disease characterized by the absence of immunity. People are susceptible, they become sick and eventually that can get the disease again. There is not immunity. In this model we have two type of interactions:

$$S + I \rightarrow 2I, \tag{2.107}$$

interaction of susceptible with infected and

$$I \rightarrow S, \tag{2.108}$$

after the disease the infected individuals come back in the susceptible compartment. This transition is spontaneous and characterized by the type of disease. For example if people on average stay sick for 3 days we can define a probability of transition of $\mu = 1/3$ because we need on average 3 time interval. Following the same arguments illustrated before we can write down the equation (for the densities):

$$\partial_t S = -\beta SI + \mu I, \tag{2.109}$$

$$\partial_t I = \beta IS - \mu I. \tag{2.110}$$

In the first rate equation we have that the first term gives the average number of S that becomes sick instead the second one gives the average number of I that recovered in the S compartment. Vice versa for I. It is important write the equation for I in a slightly different way:

$$\partial_t I = (\beta S - \mu)I, \tag{2.111}$$

in the early time $S \sim 1$ then we can write:

$$\partial_t I = (\beta - \mu)I. \tag{2.112}$$

It is clear then the behavior in early time is just function of the parameters. In fact if:

$$\beta - \mu > 0 \;\; or \;\; \frac{\beta}{\mu} > 1, \tag{2.113}$$

we have a progressive increase of the number of infected individuals and then an outbreak. Indeed if

$$\beta - \mu < 0 \;\; or \;\; \frac{\beta}{\mu} < 1, \tag{2.114}$$

we have a progressive decrease of infected individuals till $I = 0$ then the disease will not spread in all the population. The ratio β/μ is the reproductive number:

$$R_0 = \frac{\beta}{\mu}. \tag{2.115}$$

For this easy model we can calculate the analytical solution easly:

$$I(t) = \frac{\beta - \mu}{\beta + ae^{-(\beta-\mu)t}}, \tag{2.116}$$

where $a = (\frac{\beta-\mu}{I_0} - \beta)e^{(\beta-\mu)t_0}$. It is easy to see that we have two different equilibria solution for I:

$$I_\infty = 0, \;\; if \;\; R_0 < 1, \tag{2.117}$$

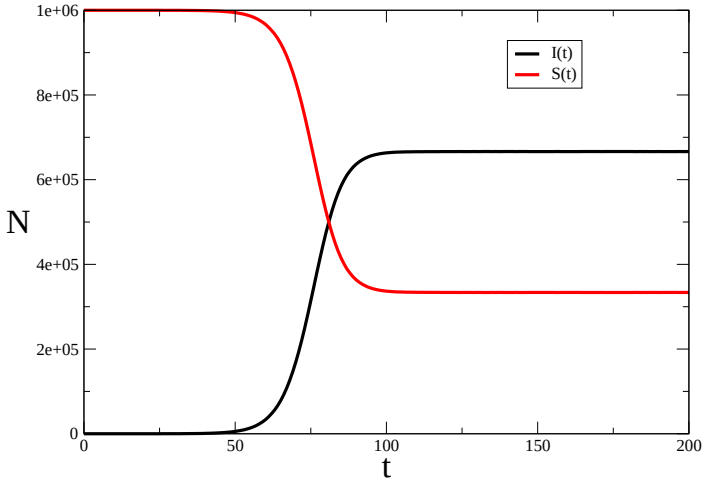

Figure 2.2: SIS model's curves (median) for $\mu = 0.1$ and $R_0 = 3$ and 1000 realizations.

and

$$I_\infty = 1 - \frac{1}{R_0}; \quad if \quad R_0 > 1. \tag{2.118}$$

It is important to stress out that for $R_0 > 1$ in this model the number of infected people reach an equilibrium value different from zero. In Figure (2.2) we show a typical example of SIS's curves I and S. We used a population of $N = 10^6$ with 1 seed ($I_0 = 1$). The profiles presented are the median of 1000 realizations. Since we modeled the simulation through a stochastic dynamic we need a sufficient number of simulations to reduce the effect of fluctuations. From the figure it is clear how the profiles reach a stationary state with both values different from zero just after 100 time steps.

2.5.3 SIR model

In this model after the disease individual get recovered (compartment R) and they are immune from the disease. The transitions are then:

$$S + I \rightarrow 2I, \tag{2.119}$$

interaction of susceptible with infected and

$$I \to R, \tag{2.120}$$

we can write down the equation (for the densities):

$$\partial_t S = -\beta S I, \tag{2.121}$$

$$\partial_t I = \beta I S - \mu I, \tag{2.122}$$

$$\partial_t R = \mu I. \tag{2.123}$$

The equation for I is the same wrote for the SIS model. Then the early time approximation and the value of R_0 is the same in both model:

$$R_0 = \frac{\beta}{\mu}. \tag{2.124}$$

There is not a general analytical solution of the equations, but we can get some interesting information from the equation.
We can consider the variation of I on respect to S:

$$\partial_S I = -1 + \frac{1}{R_0 S}, \tag{2.125}$$

that we can integrate:

$$I - I_0 = S_0 - S + \frac{1}{R_0} \ln \frac{S}{S_0}, \tag{2.126}$$

or

$$I = 1 - S + \frac{1}{R_0} \ln \frac{S}{S_0}. \tag{2.127}$$

Now a stationary point on I is reach when:

$$\partial_t I = 0 = (\beta S - \mu) I = 0, \tag{2.128}$$

this equation has two solution:

$$I = 0, \quad S = \frac{1}{R_0}. \tag{2.129}$$

Taking the second derivative of I in the stationary point we get

$$\partial_t^2 I = \beta I \partial_t S. \tag{2.130}$$

It is easy to show that $\partial_t S < 0$ then in the SIR model the stationary point can be related to $I = 0$ or a maxima and in this point the value of S is characterized by the inverse of R_0. Now using the (2.127) we can get the value of I at the peak time:

$$I_{peak} = 1 - \frac{1}{R_0} - \frac{1}{R_0} \ln R_0 S_0 = 1 - \frac{1}{R_0}(1 + \ln R_0 S_0). \tag{2.131}$$

Considering instead S and R we can write:

$$\partial_R S = -R_0 S, \tag{2.132}$$

integrating this and considering $Rt_0 = 0$ we get:

$$S = S_0 e^{-R_0 R}. \tag{2.133}$$

Since in general $0 \leq R_\infty \leq 1$ we have that after the disease a finite non zero fraction of the population is still susceptible. In Figure (2.3) we show a typical example of SIS's curves I and S. We used a population of $N = 10^6$ with 1 seed ($I_0 = 1$). The profiles presented are the median of 1000 realizations. Since we modeled the simulation through a stochastic dynamic we need a sufficient number of simulations to reduce the effect of fluctuations. From the figures it is clear that even thought the number of infected people at the peak time is the order of 210^5 more the 90 % of the population get sick. We can see this just looking at the curve of the recovered people, that is the integral of the infected one during the time. In figure (2.4) we show two different profiles the number of infected people for two different values of R_0 nominally $R_0 = 1.5, 3$. It is clear that the two realization are totally different. Big values of R_0, as we show in the early time analysis, bring a faster increase then an higher and earlier peak, as show in figure.

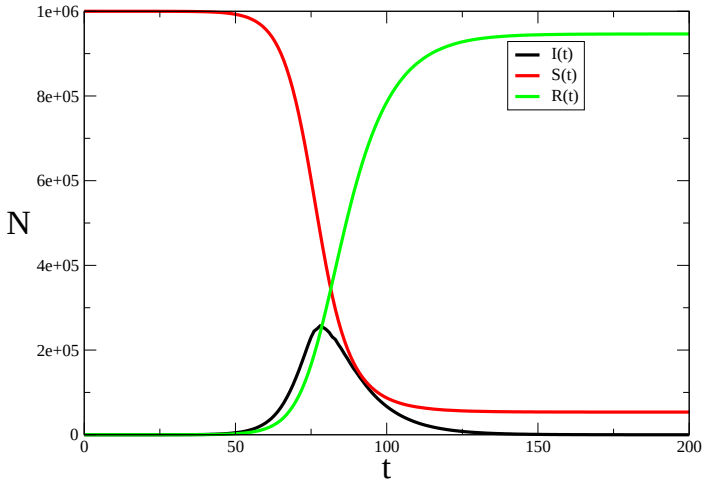

Figure 2.3: SIR model's curves (median) for $\mu = 0.1$ and $R_0 = 3$ and 1000 realizations.

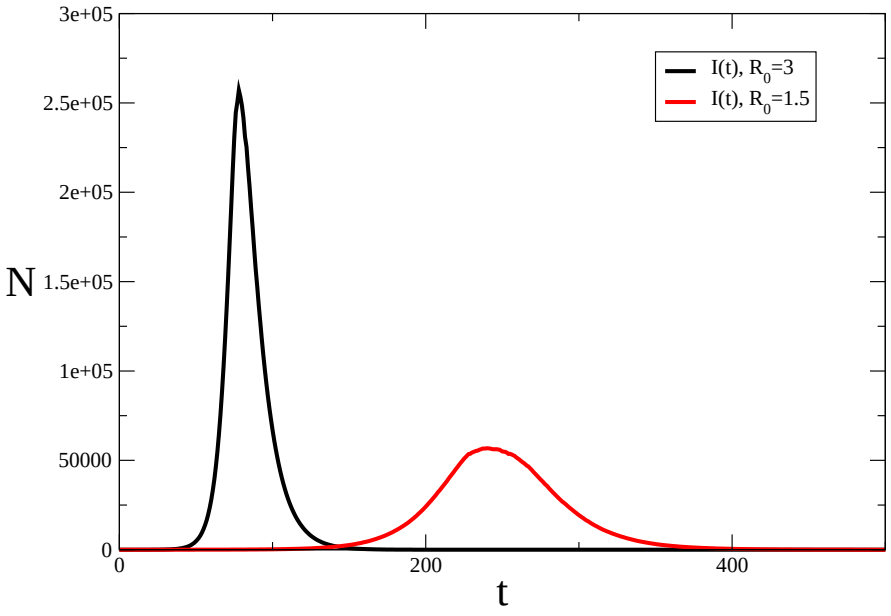

Figure 2.4: SIR model's I curves (median) for $\mu = 0.1$, $R_0 = 3$, $R_0 = 1.5$ and 1000 realizations.

3

Spectral Centrality Measures

Whether you can observe a thing or not depends
on the theory which you use. It is the theory
which decides what can be observed.
A. Einstein

Contents

In this Chapter will describe the first application of the general concepts introduced in the Chapter 2. We will described spectral centrality measures. These are usually associated to diffusion processes taking place on graphs: diffusion of importance/centrality among nodes of the networks. Not all the connections have the same importance. Not all neighbors are equivalent. In many cases the importance of a node is increased by having connections with other vertices that are themselves important. These measures take into account not just local quantity, the whole topology is considered and explored by the spread of importance among all the nodes.
We will introduce and review four centrality measures: PageRank (13), eigenvector

centrality (87) and the hub/authority scores introduced by Kleinberg for his HITS algorithm (88). These measures are usually adopted on directed graphs, we will discuss extensions to the undirected case, where applicable.

3.1 PageRank

PageRank (PR) is the prestige measure used by Google to rank Web pages. It is supposed to simulate the behavior of a user browsing the Web. Most of the times, the user visits pages just by surfing, i.e. by clicking on hyperlinks of the page he is on; otherwise, the user will jump to another page by typing its URL on the browser, or going to a bookmark, etc.. On a graph, this process can be modelled by a simple combination of a random walk with occasional jumps towards randomly selected nodes. This can be described by the simple set of implicit relations

$$p(i) = \frac{q}{n} + (1-q) \sum_{j:j\to i} \frac{p(j)}{k_{out}(j)}. \qquad i = 1, 2, \ldots, n \tag{3.1}$$

Here, n is the number of nodes of the graph, $p(i)$ is the PR-value of node i, $k_{out}(j)$ the outdegree of node j and the sum runs over the nodes pointing towards i. The *damping factor* q is a probability, that weighs the mixture between random walk and random jump. On practical applications it is usually set to small values (typically 0.15). For any $q > 0$ the process reaches stationarity, as a walker has a finite (no matter how small) probability to escape from a dangling end, whenever it lands there. When $q = 0$, the process may not be stationary and PR is ill defined. When $q = 1$, instead, the jumping process dominates and all nodes have the same PR-value $1/n$. PR goes beyond indegree: in order to have a large PR-value for a node it is important to have many neighbors pointing at a node, i.e. large indegree, but it is also important that the neighbors have large PR-values. So, if two nodes have equal indegree, the node with more "important" neighbors will have larger PR.

Solving the set of equations (3.1) is equivalent to solving the eigenvalue problem for the transition matrix $\mathcal{M}$, whose element $\mathcal{M}_{ij}$ is given by the following expression:

$$\mathcal{M}_{ij} = \frac{q}{n} + (1-q)\frac{1}{k_{out}(j)} A_{ji}. \tag{3.2}$$

PR is just the principal eigenvector of $\mathcal{M}$, and is usually determined with the power method, i.e. by repeatedly multiplying the matrix $\mathcal{M}$ by an arbitrary vector until all the entries of the resulting vector are stable. This is also the procedure we adopted to compute the eigenvectors corresponding to all centrality measures we studied.

The literature on PR is very large, because of its huge impact on Web search. In one of the first theoretical studies (89), the dependence of PR on the damping factor was investigated. In general, the attention has been mostly focused on the graph of the World Wide Web, where Web pages are nodes and the hyperlinks their connections. Comparatively little has been done to study the measure on more general classes of networks. A recent mean field study (90) has shown that the average PR value of nodes with the same indegree is a linear function of indegree in the absence of degree-degree correlations. In another study, some analytical results were found on PR distributions on special classes of graphs (91). In Section 3.4 we shall briefly resume the results of (91) and build up on them.

3.2 Eigenvector centrality

The eigenvector centrality (EV), that we briefly introduced in the Chapter 1, is also based on the principle that the importance of a node depends on the importance of its neighbors. In this case the relationship is more straightforward than for PR: the prestige x_i of node i is just proportional to the sum of the prestiges of the neighboring nodes pointing to it

$$\lambda x_i = \sum_{j:j\to i} x_j = \sum_j A_{ji} x_j = (\mathbf{A}^t \mathbf{x})_i. \tag{3.3}$$

From equation (3.3) we see that x_i is just the i-component of the eigenvector of the transpose of the adjacency matrix with eigenvalue λ. We notice that the trivial eigenvector with all components equal to zero is always a solution of equation (3.3). The true EV is then associated to the existence of non-trivial solutions of the eigenvalue problem of equation (3.3). From equation (3.3) we see that nodes with indegree zero also have zero centrality: in general, nodes pointed at by nodes with zero centrality also have zero centrality and this effect will propagate to other nodes, so that in many cases EV would not give any information about a big number of nodes. To avoid this, it is useful to make

the following modification: to each node we assign a prestige ϵ, which is independent of its relationships with the other nodes. Equation (3.3) is then modified as follows:

$$x_i = \alpha(\mathbf{A}^t\mathbf{x})_i + \epsilon. \tag{3.4}$$

The role of the parameter ϵ reminds that of the damping factor q in PR. The parameter α weighs the relative importance of the contribution of the peers versus that of the node itself. The new measure is called α-*centrality* (αEV) (87) and is the one we shall investigate in this paper. We remark that, in contrast to PR, here the solutions do not have a natural interpretation in terms of probability, so the sum of the α-centralities need not be 1. However we shall normalize the final values by dividing them by their sum, so to make them add up to 1, for practical purposes.

3.3 HITS scores

Google's PR was not the first prestige measure for Web pages based on the Web's graph representation. Shortly before the seminal paper by Brin and Page, Jon Kleinberg (88) had proposed another solution to the problem of ranking Web sites based on their importance for the users. This solution was the *HITS algorithm*, which distinguishes two types of Web pages: *hubs* and *authorities*. Let us suppose that a user submits a query through a search engine. If a page is very relevant for this query, one can reasonably expect that it will be pointed at by many other pages. However, the simple indegree would not allow to discriminate the relevant pages from other pages with similar (large) indegree. An important difference is that pages pointing to a relevant page are likely to point as well to other relevant pages, so to create a sort of bipartite structure where relevant pages (authorities) are cited by special pages/indices (hubs). Such bipartite structures allow to identify the relevant pages for the user query. Therefore one assigns two scores to a page i of the Web: the *hub score* x_i and the *authority score* y_i. Pages with high authority scores are pointed at by pages with high hub scores. In turn, a good hub points at (very) authoritative pages. This mutually reinforcing mechanism is described by the coupled relations

$$\lambda y_i = \sum_{j:j\to i} x_j = \sum_j A_{ji}x_j = (\mathbf{A}^t\mathbf{x})_i, \tag{3.5}$$

$$\mu x_i = \sum_{j:i\to j} y_j = \sum_j A_{ij}y_j = (\mathbf{A}\mathbf{y})_i, \tag{3.6}$$

which can be rewritten in the form of simple eigenvalue equations for both $\mathbf{x}$ and $\mathbf{y}$ by substitution

$$\lambda\mu x_i = (\mathbf{A}\mathbf{A}^\mathbf{t}\mathbf{x})_i. \tag{3.7}$$

$$\lambda\mu y_i = (\mathbf{A}^\mathbf{t}\mathbf{A}\mathbf{y})_i, \tag{3.8}$$

From Eqs. (3.7) and (3.8) we see that the hub and authority scores are just eigenvectors of the matrices AA^t and A^tA. We stress that both AA^t and A^tA are symmetric, whether A is symmetric or not. The scores $\mathbf{x}$ and $\mathbf{y}$ correspond to the principal eigenvectors of these matrices.

3.4 New Results

In this section we have resumed some recent results on PageRank distributions on particular types of tree-like graphs. On those graphs, the distribution of PageRank in the limit $q \to 0$ decays as a power law with exponent 2. The same is true for α-centrality, because its defining equation is formally equivalent to the equation for PageRank in the limit $q \to 0$. These results on centrality distributions are likely to be true for an extended class of graphs, where there is a flow from the outermost nodes (leaves) to a sink. We have also seen that, on any graph, in the limit $q \to 1$, the reduced PageRank of a node, i.e. the contribution of the random walk process to the measure, is simply proportional to the indegree of the node, if the nodes have (about) the same outdegree. We have studied for the first time the extension of PageRank to the case of undirected networks, finding that the reduced PageRank of a node is proportional to its degree, for large degrees, for any graph and value of q. We proposed a simple explanation of this effect based on the Central Limit Theorem, and verified numerically in several cases that the argument holds. Similarly, the reduced α-centrality of a node is also proportional to its degree, for large degrees, on any graph. With the same type of argument it is possible to show that the authority score of a node is proportional to its indegree, for large indegrees, when the outdegrees of all nodes are (approximately) the same. In the next sections we will explain all the details of these results.

PageRank

In (91) the two main limits of the PR measure, corresponding to $q \to 0$ and $q \to 1$, were investigated. Analytical results can be derived for special graphs, such as graphs grown with popular mechanisms, like preferential attachment (92). For our proofs we shall focus on the model by Dorogovtsev, Mendes and Samukhin (DMS) (62), which generates graphs with power-law degree distributions with any exponent larger than 2. As we disscused in Chapter 1 the exponent of the degree distribution of this generalization of the BA model is $\gamma = 2 + a/m$, where a is a positive constant and m is the number of links set from each new node added to the graph and the preexisting ones.

The limit $q \to 0$

We assume that q is very small. To the first order in q, and remembering that each node has outdegree 1 by construction, equation (3.1) takes the following form

$$p(i) \sim \frac{q}{n} + \sum_{j:j\to i} p(j) \qquad i = 1, 2, \ldots, n \tag{3.9}$$

which looks particularly simple, though not generally solvable. From equation (3.9) we see that the PR of a node equals a constant plus the PR of its in-neighbors. This recipe enables to calculate PR recursively on simple trees, as shown in Figure 4.2, where we focus on a subgraph of a tree. Node A is the root of the subgraph as every walk starting on any of the nodes will reach A at some stage. We call any node with this property a *predecessor* of A. The PR value of any node of the graph is determined only by its predecessors. In the case illustrated, the calculation is particularly simple: we start from the leaves of the subgraph (empty circles) whose PR is just q/n because they have no incoming links, and move towards A. For each node, we apply the relation (3.9). The final values are reported next to the nodes. From this example we can deduce a number of general properties:

- all PR values are multiples of the elementary unit q/n;
- PR increases if one moves from a node to another by following a link;
- the PR of each node i, in units of q/n, equals the number of its predecessors.

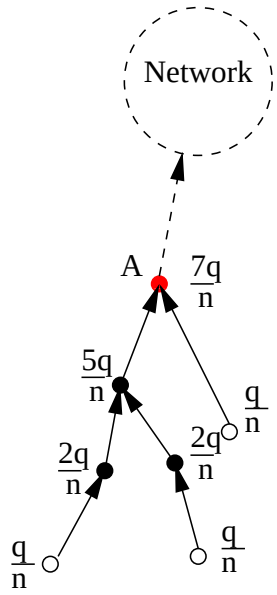

Figure 3.1: Subgraph of a tree. The PR-values of all nodes shown can be simply calculated.

Since PR takes only discrete values, in the following we shall measure it in units of q/n. We thus indicate the distribution with $P_{PR}(l)$, with $l = 1, 2, ..., n$.

In a dynamic process like network growth, it is crucial to see what happens to the PR values/distribution when a new node comes into the picture. This is shown in Figure 3.2, where a new node N is added to the network of Figure 4.2. We see that only the nodes encountered along the path from N to A, including A, are affected, while the others retain their PR values. In particular the presence of the node N determines an increase by q/n in the PR values of the affected nodes.

Now we are ready to build a master equation for the PR distribution $P_{PR}(l)$ on a DMS graph. At time n, the graph has n nodes and $n-1$ links (the root does not generate links); the PR distribution is ${P_{PR}}^n(l)$. If we add node $n+1$ we get a new distribution ${P_{PR}}^{n+1}(l)$. As we have seen above, the new node will contribute an additional q/n to the PR of the nodes in the path from $n+1$ to the root of the graph. We need to compute the balance between the nodes passing from PR $l-1$ to l and those passing from l to $l+1$. The probability Π_i^n that the PR of node i, initially equal to l, will be changed by the new node equals the probability that the link set by the new node gets attached to

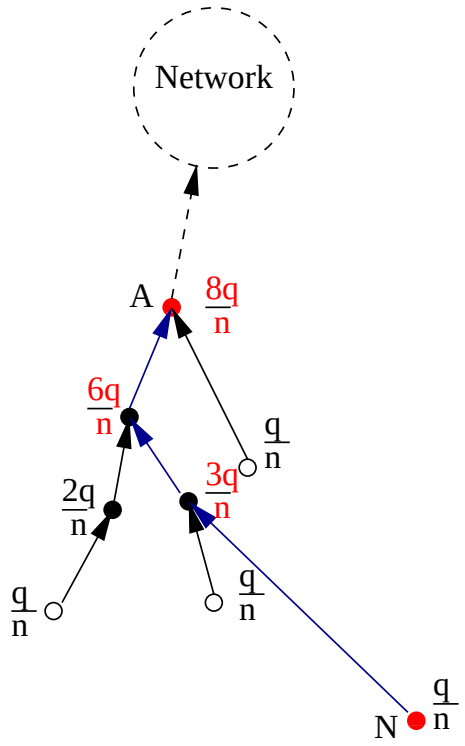

Figure 3.2: If a new node N gets attached to any node of the subgraph, it adds an equal contribution q/n to the PR of all nodes in a path from N to the root.

one of the predecessors of i (including i) and equals

$$\Pi_i^n = \sum_{j=>i} \frac{a+k_j}{\sum_{t=1}^{n}(a+k_t)} = \sum_{j=>i} \frac{a+k_j}{(a+1)n-1}, \tag{3.10}$$

where $j => i$ means that j is a predecessor of i. None of the predecessors of i, other than i can reach PR $l+1$ because of the new node, as their initial values are necessarily smaller than l. The number of predecessors of i (including i) is l and the total number of adjacent links to the predecessors is $l-1$ (one for each predecessor, except i). So,

$$\Pi_i^n = \sum_{j=>i} \frac{a+k_j}{(a+1)n-1} = \frac{(a+1)l-1}{(a+1)n-1}. \tag{3.11}$$

The number of nodes with PR l that are affected by the presence of the new node and its link is then

$$\Pi^n(l) = nP_{PR}^n(l)\Pi_i^n = \frac{(a+1)l-1}{(a+1)-1/n}P_{PR}^n(l). \tag{3.12}$$

and the master equation reads

$$(n+1)P_{PR}^{n+1}(l) - nP_{PR}^{n}(l) = \Pi^n(l-1) - \Pi^n(l). \tag{3.13}$$

quation (3.13) holds for $l > 1$. For $l = 1$ a modification is necessary, as there cannot be nodes with zero PR, so the term $\Pi^n(0)$ is not defined. However, since the new node has no incoming links, the number of nodes with PR 1 increases by 1 because of the new node, so we can write

$$(n+1)P_{PR}^{n+1}(1) - nP_{PR}^{n}(1) = 1 - \Pi^n(1). \tag{3.14}$$

The stationarity condition of Eqs. (3.13) and (3.14), in the limit of large n leads to the relations

$$P_{PR}(l) = \begin{cases} \frac{(a+1)l-a-2}{(a+1)l+a} P_{PR}(l-1), & \text{if } l > 1; \\ \frac{a+1}{2a+1}, & \text{if } l = 1. \end{cases} \tag{3.15}$$

which has the solution

$$P_{PR}(l) = \frac{a(a+1)}{[(a+1)l+a][(a+1)l-1]} \sim \frac{1}{l^2}, \text{for } l \gg 1. \tag{3.16}$$

We see that the PR distribution in the limit $q \to 0$ on a DMS tree is a power law with exponent 2, for any value of the parameter a, including the limit case $a \to \infty$, when the indegree distribution becomes exponential. This result is confirmed by numerical simulations (Figure 4.3), which also show that the hypothesis of the tree is not necessary, as long as each node has the same outdegree m.

In (91) the same result was found for other models of network growth, like Barabási-Albert preferential attachment (92) and the Copying Model (93). It is possible that this property holds for general graphs where the flows converge towards a central root (sink). Indeed, our finding agrees with the more general result on the size distribution of supercritical trees (94). Moreover, numerical studies have shown that the same behavior holds for the graph of Internet, when one considers the distribution of the size of the basin connected to a given point (95). Indeed, our calculation follows the same procedure usually adopted for the calculation of the area of basins in river networks.

The limit $q \to 1$

The case $q = 1$ is well defined, but trivial, as all nodes end up having the same PR-value $1/n$. We ask how this limit is reached. If $q \sim 1$, the contribution to PR given by the

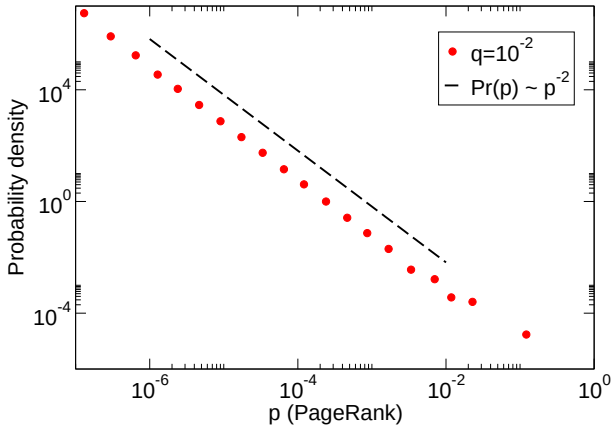

Figure 3.3: PR distribution for small q on a DMS graph with 10^6 nodes, $m = 1$ and $a = 1$. In this case the indegree distribution is a power law with exponent $\gamma = 3$.

in-neighbors of a node is very small compared to the constant term, which is close to $1/n$. In order to study the behavior of this term, we define the *reduced PageRank* $p_r(i)$ of a node i as

$$p_r(i) = p(i) - \frac{q}{n} \qquad i = 1, 2, \ldots, n. \tag{3.17}$$

We assume that all nodes have the same outdegree m. In this case, to leading order in the infinitesimal $1 - q$ equation (3.1) can be rewritten as

$$p_r(i) = \frac{q(1-q)}{mn} k_{in}(i), \qquad i = 1, 2, \ldots, n. \tag{3.18}$$

where $k_{in}(i)$ is the indegree of i. We conclude that on any graph, the reduced PR of a node in the limit $q \to 1$ is proportional to the indegree of the node, if all nodes have the same outdegree. This result has been derived independently in (96). As a consequence of equation (3.18), the distribution of the reduced PR for $q \to 1$ has the same trend as that of indegree, which can be easily verified numerically (Figure 4.4).

Extension to undirected graphs

PR can be easily extended to undirected graphs as well. The corresponding equation reads

$$p(i) = \frac{q}{n} + (1 - q) \sum_{j:j \leftrightarrow i} \frac{p(j)}{k_j}. \qquad i = 1, 2, \ldots, n \tag{3.19}$$

where now k_j is the degree of node j. For the purposes of a random walk, undirected links can be crossed in both directions, so a pure random walk now always reaches

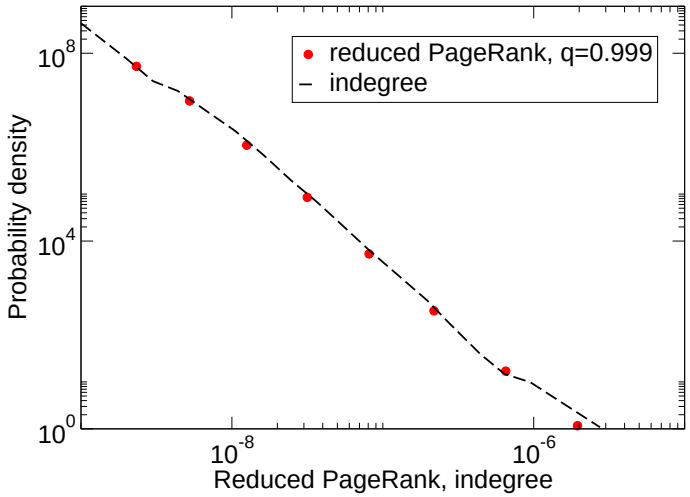

Figure 3.4: Reduced PR distribution for $q \sim 1$ on a DMS graph with 10^6 nodes, $m = 1$ and $a = 1$. The curve matches the indegree distribution.

stationarity due to the absence of dangling ends. In fact, the stationary probability of a random walk on a node of any undirected graph is simply proportional to the degree of the node (4). However, in equation (3.19) we have still the contribution of random jumping, and it turns out that the mixed process is still hard to solve. We are not aware of a general solution in this case. In the limit $q \to 0$ PR is now well behaved, and its distribution coincides with the degree distribution of the graph. In Figure 3.5 we show the distributions of reduced PR for different values of q on a DMS graph with a power law degree distribution and exponent $\gamma = 3$. The reduced PR expresses the contribution to PR given by the random walk. We see that the curves follow the decay of the degree distribution for any value of q. We have computed the reduced PR distribution on many other graphs and in all cases we found that they follow the same trend as the degree distribution. For example, in Figure 3.6 we show the comparison between reduced PR and degree for a sample of the Web link graph. Here the nodes are Web pages of the domain `.gov` and two pages are connected if there is a hyperlink from one to the other. There are $794,184$ nodes and $6,460,903$ links. The graph is directed but PR was calculated by neglecting the directedness of the links. As we can see, the decay of the distributions of reduced PR resembles that of the degree distribution. The graph at hand is not simple like the DMS networks, as it presents a large number of loops and community structure. Therefore the result is likely to be general. We can show this with

a simple argument. The general equation for reduced PR on undirected graphs is:

$$p_r(i) = \frac{(1-q)q}{n} \sum_{j:j\leftrightarrow i} \frac{1}{k_j} + (1-q) \sum_{j:j\leftrightarrow i} \frac{p_r(j)}{k_j}, \tag{3.20}$$

that we can solve formally by successive iteration, obtaining the general form

$$\begin{aligned} p_r(i) &= \frac{q}{n} \sum_s (1-q)^s \sum_{i_1} \frac{1}{k_{i_1}} \sum_{i_2} \frac{1}{k_{i_2}} \cdots \sum_{i_s} \frac{1}{k_{i_s}} \\ &= \frac{q}{n} \sum_s (1-q)^s \prod_{i_1\leftrightarrow i_2 \ldots \leftrightarrow i_s} \frac{1}{k_{i_s}}, \end{aligned} \tag{3.21}$$

where i_s indicates the neighbors of the s-shell of the node i; so, i_1 indicates the nearest neighbors of i, i_2 the next-to-nearest neighbors, and so on. The last sum in the first line of equation (3.21) is, for a given node i_{s-1}, a sum over its neighbors i_s. This sum, that we call T_{i_s}, contains k_{i_s} terms, k_{i_s} being the degree of node i_s. The sum T_{i_s} can be approximated as the product $k_{i_s}\langle 1/k\rangle_{NN}$, where $\langle 1/k\rangle_{NN}$ is the expected value of the average of $1/k$ over the neighbors of a node of the network. In general, $T_{i_s} = k_{i_s}\langle 1/k\rangle_{NN} + \eta_{i_s}$, where η_{i_s} is a random variable with mean zero. In this way, it is easy to see from equation (3.21) that, for any value of s, the product of sums reduces to $k_i\langle 1/k\rangle_{NN}$ plus the sum of many random variables like η_{i_s}. Due to the Central Limit Theorem, the latter sum, if it includes a large number of terms, yields a very small value with large probability. We can then conclude that, for k_i sufficiently large, each term of the series in equation (3.21) is proportional to k_i with good approximation, therefore $p_r(i)$ is also proportional to k_i, for any value of the damping factor q. We have verified numerically that this assertion is true for many graphs and degree distributions, without finding exceptions.

Eigenvector centrality

The defining equation (3.4) is formally analogous to equation (3.9). The only difference is that the eigenvalue α is not 1 as for PR. However, the results of Section 3.4 hold as well when the outdegree m is greater than 1 (as long as it is the same for all nodes), and in this case the sum of equation (3.9) would include a multiplicative factor $1/m$, which makes it identical to equation (3.4). We then deduce that all results found for PR in the limit $q \to 0$ hold for αEV. Here the results are more general, because we did not

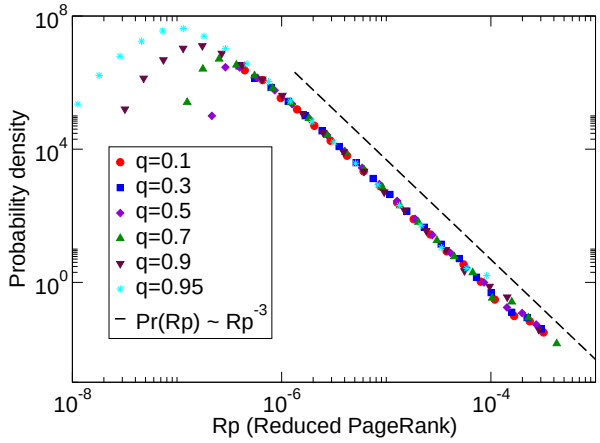

Figure 3.5: Reduced PR on undirected graphs. Variability of reduced PR distribution with q on a DMS graph with 10^6 nodes, $m = 3$ and $a = 3$. The degree distribution has a power law tail with exponent $\gamma = 3$.

need to make any approximation to get to equation (3.4) as we instead needed to derive equation (3.9). In particular, it is not necessary that ϵ be very small and the nodes need not have the same outdegree, although this is the case for the graphs we considered. We conclude that the distribution of αEV on DMS graphs has a power law tail with exponent 2 (Figure 3.7). The same holds for graphs built using preferential attachment and the Copying Model, just as it happens for PR in the limit $q \to 0$.

Extension to undirected graphs

On undirected graphs, equation (3.4) becomes

$$x_i = \alpha(\mathbf{Ax})_i + \epsilon, \tag{3.22}$$

since $A^t = A$. So, the αEV of a node is proportional to the sum of the αEV of its neighbors, modulo an additive constant ϵ. As we have done for PR, we define the *reduced α-centrality* as

$$x_i^r = x_i - \epsilon. \tag{3.23}$$

So, we can rewrite equation (3.23) as

$$x_i^r = \alpha(\mathbf{Ax^r})_i + k_i\alpha\epsilon, \tag{3.24}$$

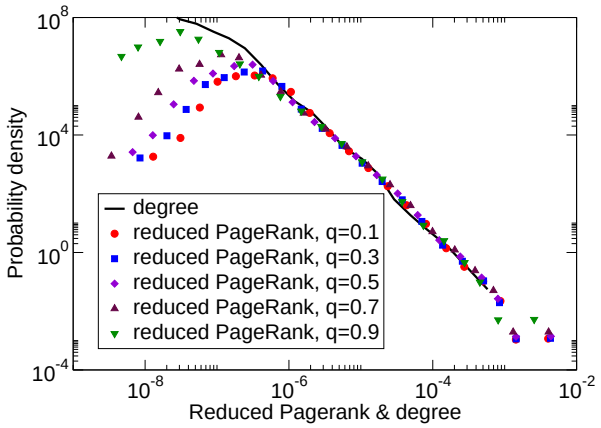

Figure 3.6: Reduced PR on undirected graphs. Variability of reduced PR distribution with q on the domain `.gov` of the World Wide Web. The degree distribution has a tail which follows fairly well a power law with exponent 2.1. To better show the agreement we have shifted the curves such that the tails overlap.

where k_i is again the degree of node i. We can apply a similar argument as in Section 3.4. The sum over the k_i neighbors of i can be approximated as $k_i\langle x^r\rangle$, where $\langle x^r\rangle$ is the average of the reduced αEV over the whole graph. The approximation is the more valid, the larger the number k_i of summands. In this way, from equation (3.24) we see that the reduced αEV of a node is proportional to its degree, if the latter is large enough. This result is independent of the specific graph we consider, and we have verified it numerically for many types of networks. In Figure 3.8 we show the distribution of reduced αEV for different choices of the parameter ϵ/α for the sample of the Web graph we analyzed in Figure 3.6. The curves closely follow the decay of the degree distribution.

HITS scores

The meaning of the eigenvalue equations (3.7) and (3.8) is quite simple. The hub score of a node is the sum of the hub scores of the in-neighbors of the out-neighbors of the node. The authority score of a node is the sum of the authority scores of the out-neighbors of the in-neighbors of the node (Figure 3.9). Let us suppose that the nodes have the same outdegree m. The authority score of a node i is given by the sum of $mk_{in}(i)$ terms, where $k_{in}(i)$ is the indegree of i. In fact, node i has $k_{in}(i)$ in-neighbors, each of them having m out-neighbors. If $k_{in}(i)$ is large, the number of summands is very large, and can be

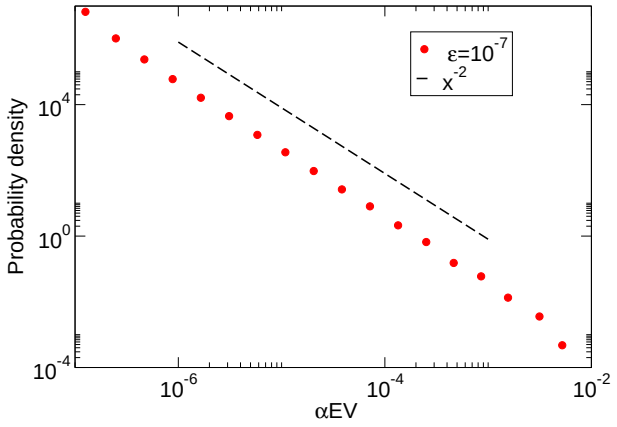

Figure 3.7: Distribution of αEV on a directed DMS graph with 10^6 nodes, $m = 1$ and $a = 1$. The dashed line indicates the predicted slope.

approximated by the average value of the authority score over the whole graph, times $mk_{in}(i)$. This approximation is the more valid, the larger m and $k_{in}(i)$. We conclude that on a directed graph with constant outdegree the distribution of the authority scores will have the same tail as the indegree distribution. This is clearly illustrated in Figure 3.10. For the hub scores it is not possible to make predictions; the sum that delivers the hub score of a node cannot be approximated through other graph variables in most cases.

The extension of the HITS scores to the case of undirected graphs is not interesting. In this case $A^t = A$, so $A^tA = AA^t = A^2$ and the hub and authority scores are identical. Moreover, they coincide with EV, as the matrices A and A^2 have the same eigenvectors.

3.5 Rankings

In the previous sections we have investigated the distributions of spectral centrality measures and their similarities. As we have mentioned centrality measures are used to rank nodes. In this section we shall compare the rankings obtained with different centrality measures. In order to compare two rankings we adopt Kendall's τ (97), a widely used index in this type of analysis. Kendall's τ ranges from 1 (perfect correlation) to -1 (perfect anticorrelation). In Table 3.1 we show the cross-comparisons between all centrality measures we discuss in this work, for a DMS directed graph. For completeness we have included the outdegree as well. As we can see, PR, αEV and the authority scores

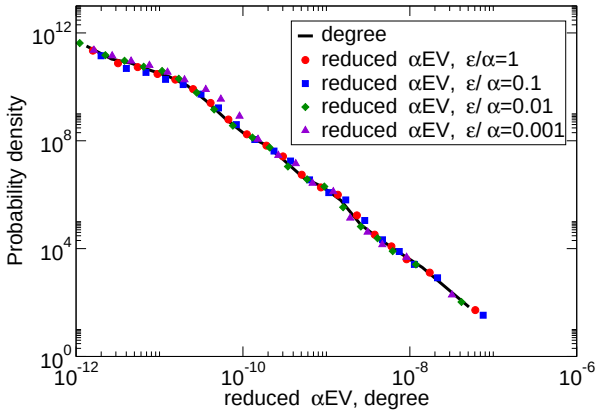

Figure 3.8: Reduced αEV on undirected graphs. Variability of reduced αEV distribution with ϵ/α on the domain `.gov` of the World Wide Web. The degree distribution has a tail which follows fairly well a power law with exponent 2.1. To better show the agreement we have shifted the curves such that the tails overlap.

are well correlated with indegree and with each other, whereas the other coefficients are small or negative; αEV has a strong correlation with outdegree as well.

DMS graphs have a fairly regular structure; we have seen that in this case the behavior of centrality measures is quite regular, and that there are simple relations between their distributions, which may be determined by simple relations between a measure and indegree at the level of the single node. Therefore, we cannot deduce general conclusions from Table 3.1 and we repeated the analysis for two real world networks: a network of political blogs and the subset of the Web link graph corresponding to the URLs of the domain `.gov`, that we have studied in the previous sections.

The first network is a citation network consisting of 1490 blogs; 758 are democratic and 732 republican. It was first studied by Adamic and Glance (98), who focused on the community structure of the graph, which matches that determined by the two political areas. The correlations now are rather weak. The small coefficients indicate that the rankings differ considerably with the measure chosen. To have an idea, in Table 3.3 we show the Top Ten blogs in the rankings obtained with all centrality measures. We see that there are clear differences between the listings.

The results are basically the same for the Web graph. Table 3.4 reports the Kendall's τ between the rankings. The values are of the same magnitude as for the network of the blogs. The Top Ten listings for the Web are shown in Table 3.5 and appear again

Measures	τ
PR-αEV	0.8192
PR-AUTH	0.5774
PR-HUBS	0.1213
PR-IN	0.6444
PR-OUT	-0.3012
αEV-AUTH	0.5788
αEV-IN	0.6487
αEV-HUBS	0.1220
αEV-OUT	0.5788
AUTH-IN	0.5458
AUTH-HUBS	0.1076
AUTH-OUT	-0.2611
HUBS-IN	0.1142
HUBS-OUT	-0.2126
IN-OUT	-0.2507

Table 3.1: Kendall's τ for each pair of centrality measures computed for a DMS directed graph, with $n = 10^6$, $m = 3$ and $a = 3$.

Measures	τ
PR-αEV	0.09
PR-AUTH	0.14
PR-HUBS	0.04
PR-IN	0.14
PR-OUT	0.02
αEV-AUTH	0.12
αEV-IN	0.07
αEV-HUBS	0.08
αEV-OUT	0.01
AUTH-IN	0.12
AUTH-HUBS	0.07
AUTH-OUT	0.01
HUBS-IN	0.02
HUBS-OUT	0.07
IN-OUT	0.07

Table 3.2: Kendall's τ for each pairs of centrality measures for the network of political blogs studied by Adamic and Glance.

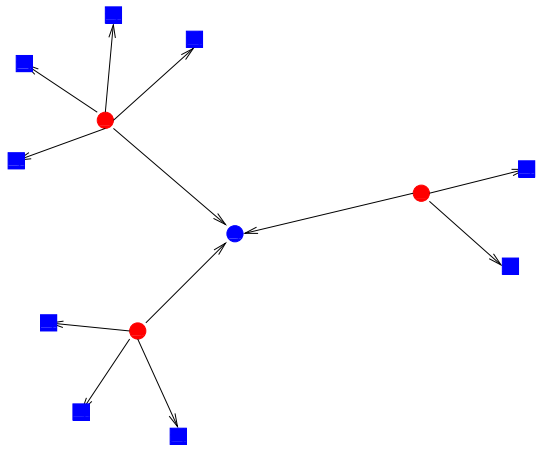

Figure 3.9: The authority score of the node in the center is proportional to the sum of the authority scores of the out-neighbors (blue squares) of the in-neighbors (red circles) of the node.

considerably different from each other.

There are often strong relations between our centrality measured and (in)degree: some relations hold on particular graphs and/or limits, others are more general. These findings imply that the measures are often strongly correlated with each other. We have indeed seen that the rankings of nodes according to the centrality measures we have considered are quite close to each other for indegree, PageRank, Eigenvector centrality and authority score on graphs built with the prescription of Dorogovtsev, Mendes and Samukhin. We have shown that these graphs have special properties, and that some measures may be correlated to each other. Instead, on real graphs, like the networks of political blogs and the sample of the Web graph we have considered, the structure is less regular and the measures are far less correlated to each other, as confirmed by the small values of the Kendall's τ for each pair of centrality measures. This means that, for practical purposes, and in spite of their similarities, spectral centrality measures look at nodes from different perspectives, and allow to diversify their roles within the network, obtaining in this way more information about the importance of nodes.

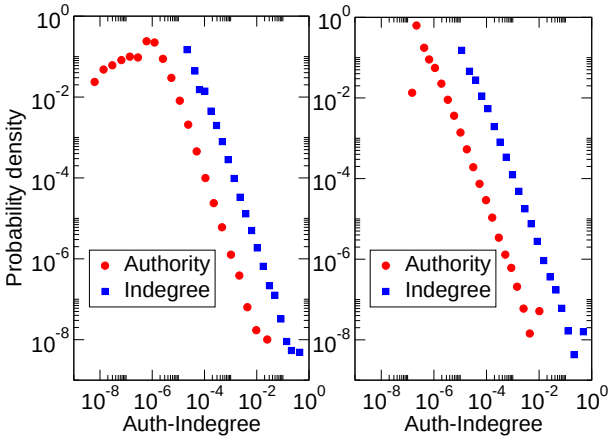

Figure 3.10: Distribution of the authority scores versus indegree distribution. (Left) DMS graph with 10^5 nodes, $m = 10$ and $a = 1$. (Right) DMS graph with 10^5 nodes, $m = 50$ and $a = 1$.

Rank	PR	αEV	Auth
1°	dailykos.com, D	atrios.blogspot.com, D	dailykos.com, D
2°	atrios.blogspot.com, D	dailykos.com, D	talkingpointsmemo.com, D
3°	instapundit.com, R	talkingpointsmemo.com, D	atrios.blogspot.com, D
4°	blogsforbush.com, R	washingtonmonthly.com, D	washingtonmonthly.com, D
5°	talkingpointsmemo.com, D	talkleft.com, D	talkleft.com, D
6°	michellemalkin.com, R	prospect.org/weblog, D	instapundit.com, R
7°	drudgereport.com, R	juancole.com, D	juancole.com, D
8°	washingtonmonthly.com, D	digbysblog.blogspot.com, D	yglesias.typepad.com/matthew, D
9°	powerlineblog.com, R	pandagon.net, D	pandagon.net, D
10°	andrewsullivan.com, R	yglesias.typepad.com/matthew, D	digbysblog.blogspot.com, D
Rank	Hubs	In	Out
1°	politicalstrategy.org, D	dailykos.com, D	blogsforbush.com, R
2°	madkane.com/notable.html, D	instapundit.com, R	newleftblogs.blogspot.com, D
3°	liberaloasis.com, D	talkingpointsmemo.com, D	politicalstrategy.org, D
4°	stagefour.typepad.com/commonprejudice, D	atrios.blogspot.com, D	madkane.com/notable.html, D
5°	bodyandsoul.typepad.com , D	drudgereport.com, R	cayankee.blogs.com, R
6°	corrente.blogspot.com, D	powerlineblog.com, R	liberaloasis.com, D
7°	aurelientt.blogspot.com, D	blogsforbush.com, R	lashawnbarber.com, D
8°	tbogg.blogspot.com, D	washingtonmonthly.com, D	gevkaffeegal.typepad.com/thealliance, R
9°	newleftblogs.blogspot.com, D	michellemalkin.com, R	presidentboxer.blogspot.com, R
10°	atrios.blogspot.com, D	truthlaidbear.com, R	corrente.blogspot.com, D

Table 3.3: Top Ten of the network of political blogs according to PR, αEV, authorities, hubs, indegree and outdegree. **D** democratic, **R**, republican.

Measures	τ
PR-αEV	0.189
PR-AUTH	0.079
PR-HUBS	0.060
PR-IN	0.155
PR-OUT	0.090
αEV-AUTH	0.081
αEV-IN	0.147
αEV-HUBS	0.074
αEV-OUT	0.086
AUTH-IN	0.046
AUTH-HUBS	0.109
AUTH-OUT	0.072
HUBS-IN	0.003
HUBS-OUT	0.056
IN-OUT	0.081

Table 3.4: Kendall's τ for each pairs of centrality measures for the domain `.gov` of the Web.

Rank	PR	αEV
1°	www.usgs.gov	polar.wwb.noaa.gov/waves/main_int.js
2°	www.nws.noaa.gov	polar.wwb.noaa.gov/waves/welcome.html
3°	www.naca.larc.nasa.gov/readme.html	polar.wwb.noaa.gov/waves/main_table.html
4°	www.usda.gov	polar.wwb.noaa.gov/waves/products.html
5°	www.nws.noaa.gov/disclaimer.html	polar.wwb.noaa.gov/waves/main_int.html
6°	www.ar.inel.gov/home.htm	www.nws.noaa.gov/disclaimer1.html
7°	www.4woman.gov/search/search.cfm	www.nws.noaa.gov
8°	www.nws.noaa.gov/feedback.shtml	polar.wwb.noaa.gov/waves/references.htm
9°	www.access.wa.gov	polar.wwb.noaa.gov/waves/validation.htm
10°	www.usinfo.state.gov/products/pdq/pdq.htm	polar.wwb.noaa.gov/waves/valid_wna.html
Rank	Auth	In
1°	www.srh.noaa.gov/oun/cgi-bin/wxclick.pl?county=oklahoma	www.usgs.gov
2°	www.srh.noaa.gov/oun/cgi-bin/wxclick.pl?county=cleveland	www.cdc.gov
3°	www.srh.noaa.gov/oun/cgi-bin/wxclick.pl?county=kiowa	www.usda.gov
4°	www.nws.noaa.gov	www.doi.gov
5°	www.srh.noaa.gov/oun/cgi-bin/wxclick.pl?county=logan	www.nws.noaa.gov
6°	www.srh.noaa.gov/oun/cgi-bin/wxclick.pl?county=payne	www.usgs.gov/disclaimer.html
7°	www.srh.noaa.gov/oun/cgi-bin/wxclick.pl?county=knox	www.usda.gov/news/privacy.htm
8°	weather.noaa.gov	www.abag.ca.gov
9°	weather.noaa.gov/weather/ok_cc_us.html	www.ars.usda.gov/nodisc.html
10°	www.crh.noaa.gov/ddc	www.ars.usda.gov/comm.htm

Table 3.5: Top ten of the web domain `.gov` according to PR, αEV, authorities and indegree.

4

PageRank Localization

Physics is like sex: sure, it may give some practical results, but that's not why we do it.

R.Feynman

Contents

In this Chapter, we focus on a spectral centrality measure particularly important for the applications to information technology: the PageRank (PR) p_i of the vertex i. In particular, we propose a new interpretation for this quantity that can in principle change the way in which this quantity is computed and analyzed in its dynamical evolution. We propose that computing the PR can be recast in the form of several well known problems in physics, from the charge-distribution in an inhomogeneous medium to the wave-localization phenomena in quantum-physics, at the cost of replacing the continuous space by a discretized and oriented graph (99).

As we discussed in the Chapter 3 PR is defined as the stationary distribution of a discrete-time, finite-state Markov chain given by a random walk on the graph(100). In other words the PR of a vertex can be defined as the time spent by a random surfer on

it. The PR is defined as:

$$p(i) = \frac{1-\alpha}{n} + \alpha \sum_{j:j\to i} \frac{p(j)}{k_{out}(j)} \qquad i = 1, 2, \ldots, n \tag{4.1}$$

Or introducing the vector $\mathbf{p}$ can be seen as the solution of the equation

$$\mathbf{p} = \left[\alpha \left((\mathbf{K}^O)^{-1}\mathbf{A}^T\right) + (1-\alpha)\mathbf{E}\right] \mathbf{p} \tag{4.2}$$

where E is a matrix whose elements are $1/N$, α is the damping factor weights the two contributions (in this Chapter we are using a complementary notation on respect to the definition used in the Chapter 3. In particular $q = 1 - \alpha$) and the correct stochastic matrix of the process is given by $\left[\alpha \left((\mathbf{K}^O)^{-1}\mathbf{A}^T\right) + (1-\alpha)\mathbf{E}\right]$.

4.1 Novel formalization

Here we write equation (4.1) in different way in order to get some more physical insight. The first algebraic passage is to introduce, in the strongly connected component (SCC) of the graph, the variable $\psi_i = p_i/k_i^O$ and then to add and subtract the term $\alpha\psi_i k_i^I/k_i^O$ to equation (4.1). In this way we obtain

$$\psi_i = \frac{\alpha}{k_i^O}\left(\sum_{j\to i}\psi_j - k_i^I\psi_i\right) + \alpha\frac{k_i^I}{k_i^O}\psi_i + \frac{1-\alpha}{k_i^O}\frac{1}{N}, \tag{4.3}$$

and after some rearrangements

$$-\left[\sum_{j\to i}\psi_j - k_i^I\psi_i\right] + \left[\frac{k_i^O - \alpha k_i^I}{\alpha}\psi_i\right] = \frac{1-\alpha}{\alpha}\frac{1}{N}. \tag{4.4}$$

We can now spot in (4.4) the two distinct terms in square brackets. As discussed below, the first one can be put in relation with a discretized graph operator, while the second one can be considered as a potential function $\mathbf{V}$ such that equation (4.4) can be rewritten as

$$\left[-\nabla^{\mathbf{2}}_{\mathbf{D}} + \mathbf{V}\right]\boldsymbol{\psi} = \mathbf{F} \tag{4.5}$$

Here, $\mathbf{F}$ is a vector with constant entries $F_i = (1-\alpha)/\alpha N$, the components of $\mathbf{V}$ are given by $(k_i^0 - \alpha k_i^I)/\alpha$ and the operator $\nabla^2_D\psi_i = \sum_{j\to i}\psi_j - k_i^I\psi_i$ can be *defined* as the

discretized Laplacian on an oriented graph. Indeed, the discretized Laplacian operator (acting on a test vector ϕ_i) on a regular lattice is given by

$$(\left[\nabla^2\right]\phi)_i \equiv \sum_{<j,i>} \phi_j - k\phi_i \tag{4.6}$$

where the sum runs on all the k neighbors j of site i. The Laplacian of ϕ in a point is given by the sum of the values on all the neighbors minus k times (where k is the degree or connectivity of the lattice) the value in the site considered.

The operator ∇^2_D in equation (4.5) is the directed counterpart of the discretized Laplacian (hence its subscript "D"). In the case of the WWW and, more in general, for any directed graph, the edges can be travelled only in one direction. This means that the lattice is not "reciprocal", where with reciprocity we intend the property in a graph to reach vertex B starting from A, if A can be reached from B (see also Ref.(101)). In particular, if for every edge in a graph from i to j there is also the symmetric edge j to i, the *reciprocity* is 100%. More generally the reciprocity is given by the fraction of symmetric edges in the graph.

4.2 Directed laplacian and localization

To our knowledge the properties of the Laplacian on these directed lattices have never been considered before. The operator ∇^2 and its directed counterpart ∇^2_D will, in general, be very different; however we expect this to be dependent on the reciprocity. We tested the solution of the Laplace equation $\nabla^2\boldsymbol{\psi} = 0$ against the correspondent directed Laplace equation $\nabla^2_D\boldsymbol{\psi} = 0$ in a series of lattices (both regular as the simple cube and some realizations of Barabási-Albert network) starting from a completely reciprocal case and deleting randomly some of the connections. Generally, provided that the proportion of reciprocal links is above the percolation threshold for the lattice considered(102), the statistical behavior of the directed Laplacian is the same of the non-directed one (See Figure 4.1).

In the particular case in which the network is reciprocal ($\mathbf{V} = 0$) no trapping states are present and we can therefore consider $\alpha = 1$ (89). Equation (4.5) becomes then a Laplace equation whose solution, $\boldsymbol{\psi}$, is given by a constant function. We find that in this case the PR is proportional to the degree of the page(91). This limit case in which a constant distribution of $\mathbf{V}$ gives a trivial distribution of $\boldsymbol{\psi}$ suggest that the term with $\mathbf{V}$ in (4.5)

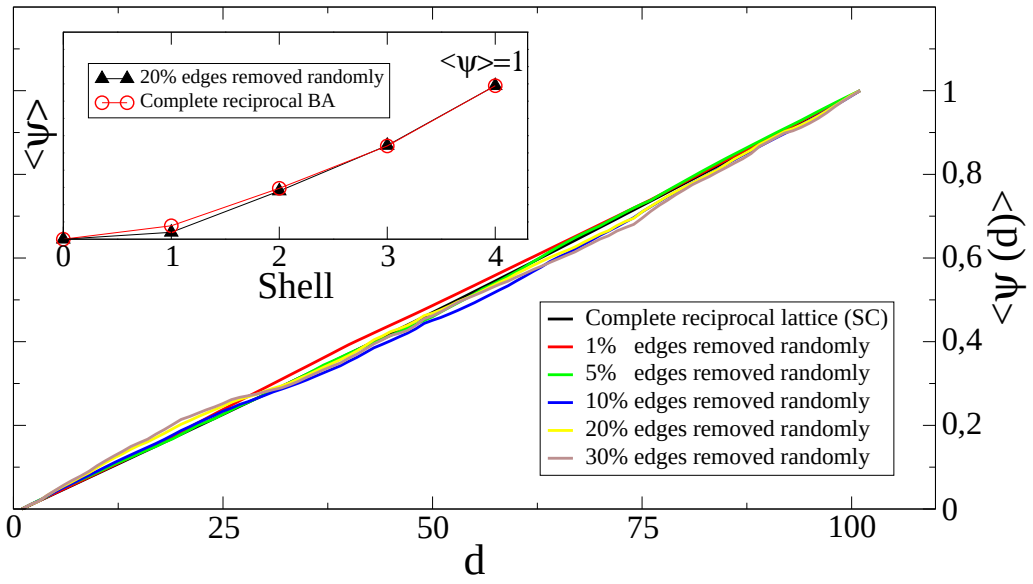

Figure 4.1: Plot of ψ, obeying the equation $\nabla^2\psi = 0$ (complete reciprocal lattice and complete reciprocal Barabási-Albert graph) and $\nabla^2_D\psi = 0$ (all the other curves) in a 2-d simple cube lattice and in a Barabási-Albert network. In the simple cubic case the upper and lower layer are kept at the fixed value of $\psi = 0$ and $\psi = 1$ respectively. The average value increases from 0 to 1, even when the reciprocity in the links between pairs of nodes is broken up to 30% of the completely reciprocal case. In the inset the plot of the same quantity for a Barabási-Albert model where the same boundary conditions of $\psi = 1$ and $\psi = 0$ are applied to the leaves and to the core of the structure respectively. Here four shells are considered starting from a node in the core and the behavior is the same even for a 20% removal of reciprocal links

can be considered similar to that of a potential, and we can expect that the PR will be localized around the minima of such a potential. To test this hypothesis we considered a snapshot of the real WWW, collected by Dipartimento di Scienze dell'Informazione (DSI), Univ. degli Studi di Milano and consisting of the 786.049 pages of the .eu domain connected by 18.120.539 edges (103). In Figure 4.2 we show a 3D representation of the potential and the corresponding value of PR visualized as follows: given one reference node (the node with the highest PR value in this case) that is put in the center of a conventional (x, y) plane, we arrange in circular shells the first, second, third... n-th neighbors. For each cell the neighbors are settled as equispaced points in a circle and the corresponding Potential and PR are given in the z axis. The resulting 3D plots are

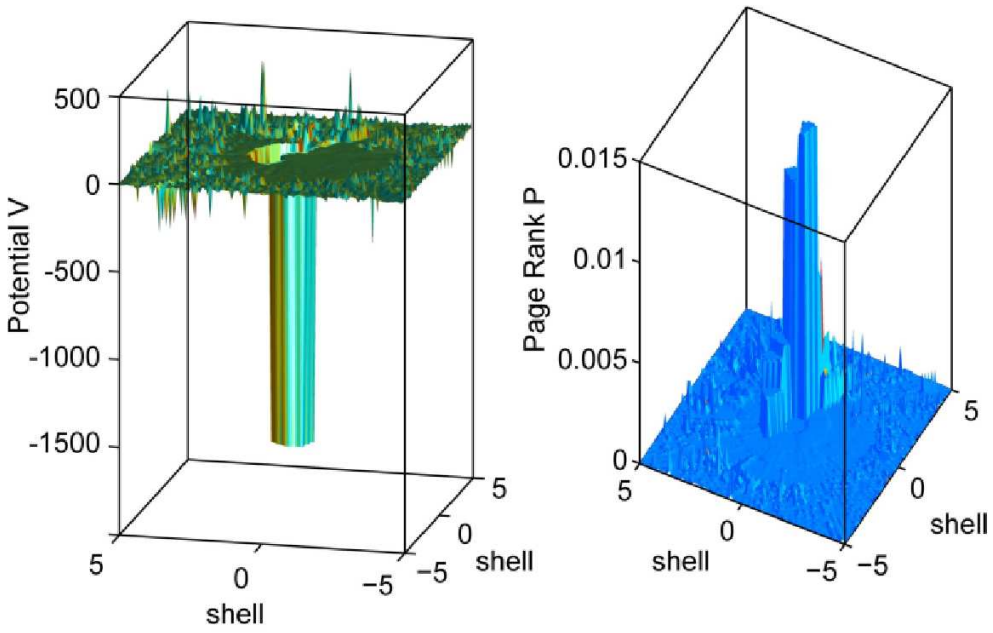

Figure 4.2: (Color online) 3D plot of potential $\mathbf{V}$ and the corresponding PR measured along concentric shells around the vertex with the highest value of PR.

given in Figure 4.2, and the highest PR value node is such that the potential display a pronounced minimum, and correspondingly the PR is localized. In Figure 4.3 we show the PR after averaging over shell nodes as linear plot.

Within this framework, the higher scores of the PR will be **localized** in the potential wells, and correspondingly nodes associated to peaks of the potential ("repulsive" regions) will display a low PR, as shown in Figure 4.4. This clarifies the role of the potential, in the real WWW there is no complete reciprocity and the difference between outdegree and indegree of a page plays the role of a topological disorder. It is important to stress that, since the Web has not a simple topology, the fact that a page is a minimum or a maximum of the potential is only evident when plotting the values at the nearest neighbors in the network. By exploiting topological information in the potential $\mathbf{V}$, one is able to gain information about the PR spatial distribution. In particular by using the value of the potential as a rule of thumb to determine PR value we are able to spot in the above data 61 among the top 100 values of PR. To improve this result (without considering the general solution for ψ), one would have to take into account gradients

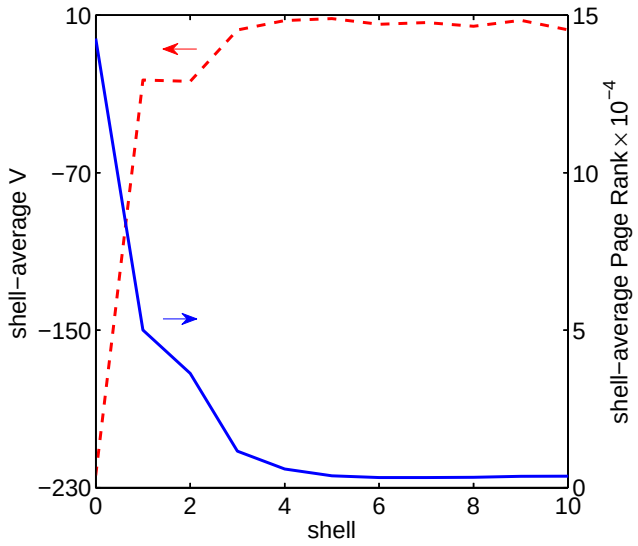

Figure 4.3: (Color online) Shell average potential (red, left scale, dashed line) and shell **PR P** (blue, right scale, continuous line) as obtained in the presence of an *hub*, i.e. a site with large in-degree, small out-degree and therefore very low potential. These quantities are computed over concentric shells of neighbors. The shell average is obtained by averaging over nodes on the same shell.

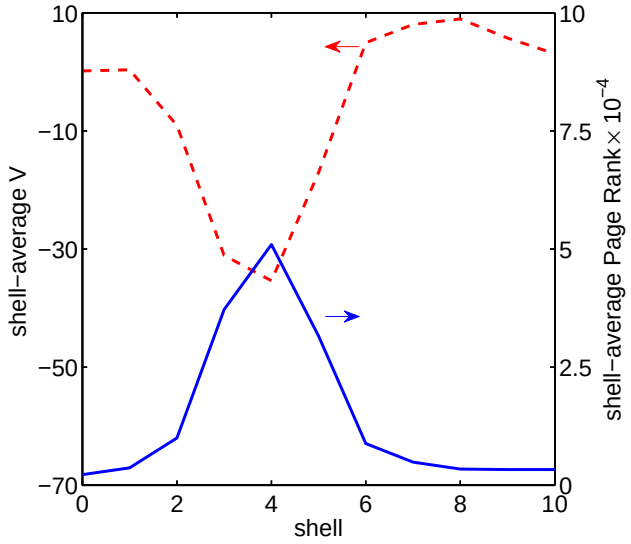

Figure 4.4: (Color online) As in the above figure, both the shell average potential (red, left scale, dashed line) the and **PR P** (blue, right scale, continuous line) in the presence of the maximum of the potential, produced by a node with small in-degree and large out-degree.

and expansions around the local **V** minima. Equation (4.5) can then be interpreted in various ways, from a Poisson equation in a disordered medium to an inhomogeneous Helmholtz equation(104). We decided to exploit the similarity with a time independent Schrödinger equation (with the addition of a constant term on the r.h.s.) because, in our opinion, this clarifies in a particularly clear way the role of the potential **V**. In this perspective, the r.h.s. of (4.5) plays the rule of a stochastic source as often encountered, e.g., in the c-number representation of quantum field-theories (105).

4.3 Alternative method to evaluate the PageRank

This interpretation suggests a relatively simple way to compute the whole distribution of the PR. In principle once the matrices of the Laplacian operator and the potential operator are known, the ψ (and henceforth the set of PR values) could be computed by inverting these operators. This simple operation is unfeasible when the size of the matrix is of the order of tens of billions of pages as in the WWW. Here we adopt a different approach based on a matrix expansion that can be also extended to study the

time evolution. The idea is to rewrite the above equation by using the common Taylor expansion, while starting from the equation (4.5):

$$\boldsymbol{\psi} = \left(\mathbf{I} - \mathbf{V}^{-1}\boldsymbol{\nabla}_{\boldsymbol{D}}^{\mathbf{2}}\right)^{-1}\mathbf{V}^{-1}\mathbf{F}. \tag{4.7}$$

We now expand the expression in brackets by writing

$$\left(\mathbf{I} - \mathbf{V}^{-1}\boldsymbol{\nabla}_{\boldsymbol{D}}^{\mathbf{2}}\right)^{-1} = \sum_{n=0}^{\infty}\left(\mathbf{V}^{-1}\boldsymbol{\nabla}_{\boldsymbol{D}}^{\mathbf{2}}\right)^{n} \tag{4.8}$$

provided all the eigenvalues λ_h of $\left(\mathbf{V}^{-1}\boldsymbol{\nabla}_{\boldsymbol{D}}^{\mathbf{2}}\right)$ have $|\lambda_h| < 1$.
This allows to invert only the diagonal matrix $\mathbf{V}$ (that can be done easily by taking the inverse of the elements on the diagonal). The expression above can be rewritten as

$$\boldsymbol{\psi} = \left(\mathbf{k}^{O} - \alpha\mathbf{A}^{T}\right)^{-1}\mathbf{F}' = (\mathbf{I} - \alpha\boldsymbol{B})^{-1}\left(\mathbf{k}^{O}\right)^{-1}\mathbf{F}' \tag{4.9}$$

where $\mathbf{F}' = \alpha\mathbf{F}$, $\mathbf{k}^{O}$ is a matrix whose elements are all zero apart on the diagonal where they are given by the outdegree of vertices, $\mathbf{A}^{T}$ is the transpose of the adjacent matrix and $\mathbf{B} = \left(\mathbf{k}^{O}\right)^{-1}\mathbf{A}^{T}$.
Equation (4.9) closely resembles the original equation for PR, with the important caveat that we are now working with a wave function $\boldsymbol{\psi}$. In this case, the expansion:

$$(\mathbf{I} - \alpha\mathbf{B})^{-1} = \sum_{n=0}^{\infty}(\alpha\mathbf{B})^{n} \tag{4.10}$$

converges and we can calculate with the desired precision $\boldsymbol{\psi}$ and so the associated PR. The results of this matrix expansion are in good agreement with the solution obtained by traditional methods. One can increase as desired the order of the expansion with a computational cost that increases only linearly with the order.

5

Behavioral Changes

It's not that I'm afraid to die, I just don't want to be there when it happens.
W. Allen

Contents

Human behavior has long been recognized as one of the key points in understanding epidemic spreading (106), leading to a concerted effort to include social complexity in epidemiological models. Age structure (107), air line travel (108; 109) and commuting are now incorporated in all realistic models (21). However, much remains to be done. The recent H1N1 pandemic has demonstrated the feasibility and reliability of epidemic forecasting in real time (16; 110), but it has also brought to light the limitations underlying the state of the art of epidemic models (17). In particular, it has become clear that societal reactions can have an important impact on epidemic spreading (111). These reactions can be classified into different classes. In the first, changes are imposed by authorities through the closure of schools, churches, public offices and bans of public gatherings (112; 113). In the second, individuals decide to modify their behavior due to concern about a disease, by avoiding social contacts with infected individuals and

crowded spaces, reducing traveling or preventing children from attending school. In both cases we will have a modification of the spreading process due to the reduction of contacts in the population. In general, the result of these measures is very important. A reduction of the epidemic outbreak and a delay of the epidemic peaks are possible outcome. For these reasons social distancing policies are crucial measures during serious epidemic spreading.

Many studies have been done in order to evaluate the impact and role of organized public health measures in real epidemics (112; 114; 115). Instead just a few, recent, attempts have considered spontaneous social distancing phenomena. In some approaches individual behaviors were modeled by introducing different contact rates *(normal* or *altered)* in response to the state of disease (116; 117), in others new compartments representing individual states or levels of self-imposed isolation were proposed (118) while in others the spreading of awareness was coupled with the disease (119). However, there is no consensus on how spontaneous social distancing is related to the perceived current state of the epidemic. The definition of a general model is still an open issue.

In this study we propose a general framework to model the spreading of awareness and social distancing in a single population. We modify the classical SIR model (79) by introducing a new compartment, S^F, that represents susceptible people aware of an infectious disease. These people decide to reduce their number of contacts as a way of trying to reduce the likelihood of becoming infected. We modeled the spread of epidemic awareness within the population considering different mechanisms. We related the awareness to the state of the epidemic at a given time, to the number of scared people (through "fear contagion" (120)) and to the information that spreads by the media (121). In this Chapter we present a complete survey of these different processes, their implementation and the analysis of their main features.

5.1 Fear of the sick

The first model we considered, is a generalization of the SIR model, on a single population, that includes a new compartment of susceptibles: S^F [1]. Individuals in this compartment are more careful about their contacts, thus reducing the contagion rate

[1] where F is taken to refer to fear

$\beta \rightarrow r_\beta \beta$ with $(0 \leq r_\beta < 1)$. *Normal* susceptible people reach the S^F compartment (become feared) after interacting with infected people in compartment I. In this case, fear is generated by the presence of infected persons in the community. This process can be considered as a parallel contagion process. By analogy we defined the reproductive number for the fear as:

$$R_F = \frac{\beta_F}{\mu_F}. \tag{5.1}$$

People can recover from *fear* and return into the susceptible compartment by interacting with recovered people, R, and other susceptibles, S, with a constant rate μ_F. People stop being afraid after seeing that not many people are affected by the disease and that the ones that were infected are now recovered. The full epidemic model is then described by the following set of equations:

$$\begin{aligned}
d_t S(t) &= -\beta S(t)\frac{I(t)}{N} - \beta_F S(t)\frac{I(t)}{N} + \mu_F S^F(t)\left[\frac{S(t)+R(t)}{N}\right], \\
d_t S^F(t) &= -r_\beta \beta S^F(t)\frac{I(t)}{N} + \beta_F S(t)\frac{I(t)}{N} - \mu_F S^F(t)\left[\frac{S(t)+R(t)}{N}\right], \\
d_t I(t) &= -\mu I(t) + \beta S(t)\frac{I(t)}{N} + r_\beta \beta S^F(t)\frac{I(t)}{N}, \\
d_t R(t) &= \mu I(t).
\end{aligned}$$

It is important to stress that we considered a process in which:

$$\sum_i d_t X_i(t) = 0 \quad for \ \forall \ t \ \ and \ \ X_i \in \left[S, S^F, I, R\right], \tag{5.2}$$

meaning that the total number of individuals in the population does not change. In diseases like flu, the time scale of the spreading is very small on respect to the average life time of a person. This allowed us to ignore birth or death processes. The dynamics take place with a fix number of individuals. The flows between compartments balance each others. For each negative term there is another one, equal, but with a positive sign. To explain the equations we can just consider negative terms. In particular: in the first equation in the (5.2) the first term takes into account individuals in the susceptible compartment S that interacting with infected individual become sick:

$$S + I \xrightarrow{\beta} 2I. \tag{5.3}$$

The second term takes into account individuals in the susceptible compartment S that interacting with infected individuals become scared by disease:

$$S + I \xrightarrow{\beta_F} S^F + I. \tag{5.4}$$

The first term of the second equation takes into account individuals in the compartment S^F that interacting with infected individuals get sick:

$$S^F + I \xrightarrow{r_\beta \beta} 2I. \tag{5.5}$$

This happen with a rate $r_\beta \beta < \beta$ because, as we said, people aware of the disease reduce their contacts. The last term in the second equation takes into account people in the compartment S^F that interacting with healthy individuals, S, and recovered ones, R, stop to be aware and move back in the compartment S:

$$S^F + S \xrightarrow{\mu_F} 2S, \tag{5.6}$$

and

$$S^F + R \xrightarrow{\mu_F} S + R. \tag{5.7}$$

The first term in the third equation takes into account the spontaneous recovery of sick individuals:

$$I \xrightarrow{\mu} R. \tag{5.8}$$

When the disease spreads much faster than public awareness, the model reduces to the classical SIR, with basic reproductive number, $R_0 = \beta/\mu$. In this limit, the early time of the compartment S^F is given by (assuming $S^F_{t=0} \equiv 0$):

$$S^F(t) \sim \frac{\beta_F}{\mu(R_0 - 1) + \mu_F} \left(e^{\mu(R_0-1)t} - e^{-\mu_F t} \right). \tag{5.9}$$

It is clear a transition between two regimes. For

$$\mu(R_0 - 1) > \mu_F, \tag{5.10}$$

the rate of increase of the fear is governed by R_0, otherwise fear dies out: the rate of fear production it is not enough to sustain it. However, when panic spreads faster than the disease, $\left(R_F \gg R_0^{SIR}\right)$ everyone quickly becomes scared and our model reduces to an SIR model with a reduced reproductive rate $R_0^F = r_\beta \beta/\mu$, dominated by the characteristics of the S^F compartment. We explored numerically the intermediate regime between

these two limits. For small values of R_F the presence of fear does not significantly affect the timing of the disease, as showed in Figure (5.1). It simply produces a mild reduction on epidemic size. Around the 20% for $r_\beta = 0$ and $R_F = 5$ (see Figure (5.2)).
Increasing the value of R_F results in two different scenarios:

1. $r_\beta \beta/\mu > 1$ the reduction of epidemic size is bounded to the value of an SIR model with $\beta \to \beta r_\beta$;

2. $r_\beta \beta/\mu < 1$ fear completely stops the progression of the disease.

After the end of the epidemic, the system enters in the so called disease-free equilibrium. In the phase space this is describe by:

$$(S_\infty, S^F_\infty, I_\infty, R_\infty) = (1 - R_\infty, 0, 0, R_\infty). \tag{5.11}$$

From the equations above, it is easy to show that fear disappears exponentially:

$$S^F(t) \sim e^{-\mu_F t}. \tag{5.12}$$

There is not possibility of an endemic state of fear. Fear can only be produced by the presence of infected people, as soon the infection dies, scared people can recover from fear by interacting with all the susceptible and recovered becoming susceptible themselves.

5.2 Self-reinforcing fear

Until now we have not considered the possibility that one might enter in the compartment S^F simply by interacting with people already in this compartment, fear generating fear. Mathematically, this process is modeled by treating fear as a thought contagion process: people become scared by interacting with scared people:

$$S + S^F \xrightarrow{\beta_F \alpha} 2S^F. \tag{5.13}$$

A new parameter, $\alpha \geq 0$, is necessary to distinguish between infection due to contacts with infected and feared people. Assuming that people who contact infected people are more likely to be aware of the disease than people who interact with feared individuals,

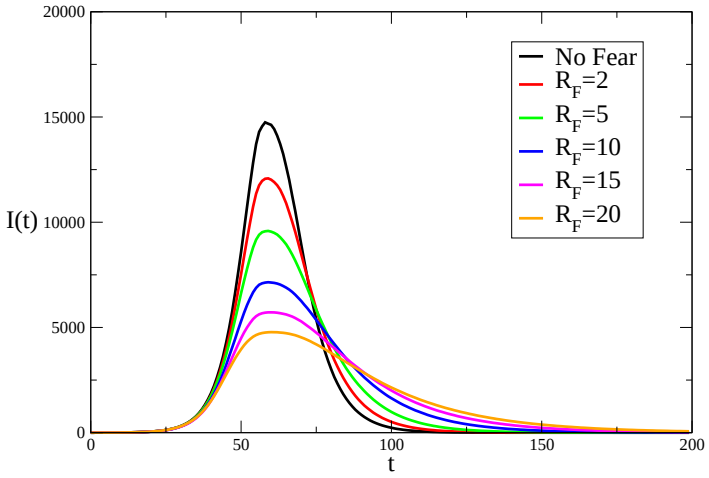

Figure 5.1: For $\mu_F = 0.5$ and $r_\beta = 0.5$ median of $I(t)$ profiles for different values of R_F are plotted on respect to the baseline without *fear*

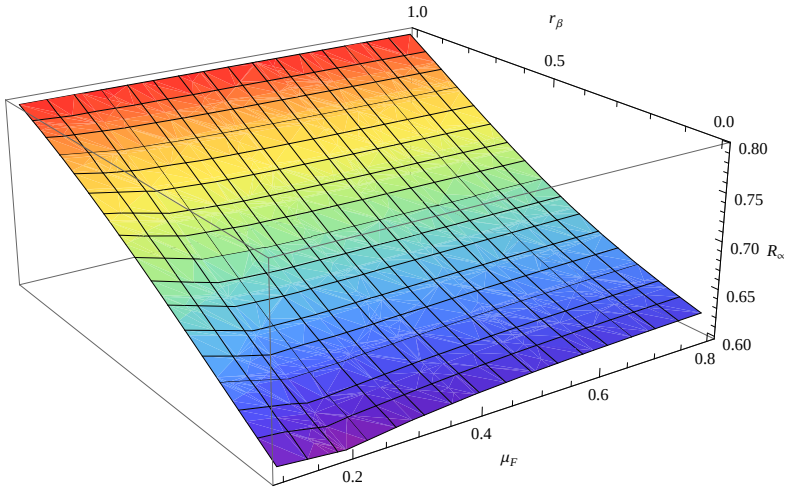

Figure 5.2: Epidemic size R_∞ for different values of μ_F, r_β and $R_F = 5$

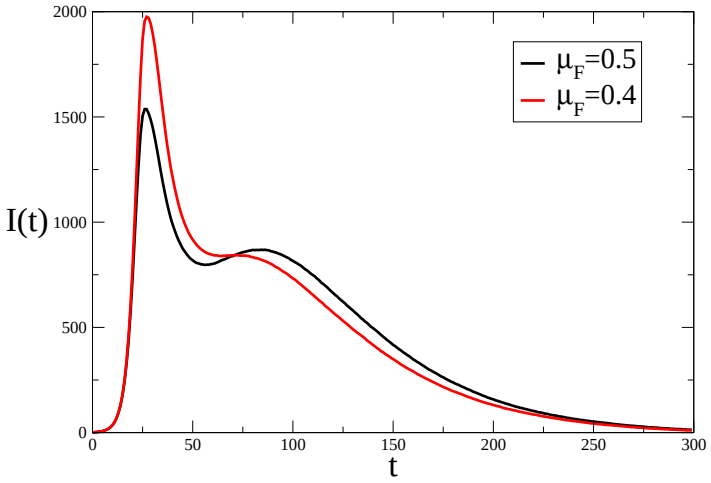

Figure 5.3: Median of $I(t)$ profiles for $\mu_F = 0.5$, $r_\beta = 0.42$, $\alpha = 0.05$, $R_0 = 2$, $\mu = 0.4$ and $R_F = 1.2$

we can set: $0 < \alpha < 1$. The equations that describe our model are then:

$$\begin{aligned}
d_t S(t) &= -\beta S(t)\frac{I(t)}{N} - \beta_F S(t)\left[\frac{I(t)+\alpha S^F(t)}{N}\right] + \mu_F S^F(t)\left[\frac{S(t)+R(t)}{N}\right], \\
d_t S^F(t) &= -r_\beta \beta S^F(t)\frac{I(t)}{N} + \beta_F S(t)\left[\frac{I(t)+\alpha S^F(t)}{N}\right] - \mu_F S^F(t)\left[\frac{S(t)+R(t)}{N}\right], \\
d_t I(t) &= -\mu I(t) + \beta S(t)\frac{I(t)}{N} + r_\beta \beta S^F(t)\frac{I(t)}{N}, \\
d_t R(t) &= \mu I(t).
\end{aligned}$$

As before, if we assume that the disease spreads faster than the awareness the reproductive ratio is $R_0 = R_0^{SIR}$. The early time evolution of the S^F compartment is given by:

$$S^F(t) \sim \frac{\beta_F}{\mu(R_0-1) - \mu_F(R_F-1)}\left(e^{\mu(R_0-1)t} - e^{\mu_F(R_F-1)t}\right), \tag{5.14}$$

where we defined

$$R_F = \alpha\beta_F/\mu_F. \tag{5.15}$$

If $\mu(R_0-1) > \mu_F(R_F-1)$ the rate of increase of fear is dominated by the epidemic, otherwise, it is dominated by its own thought contagion process.

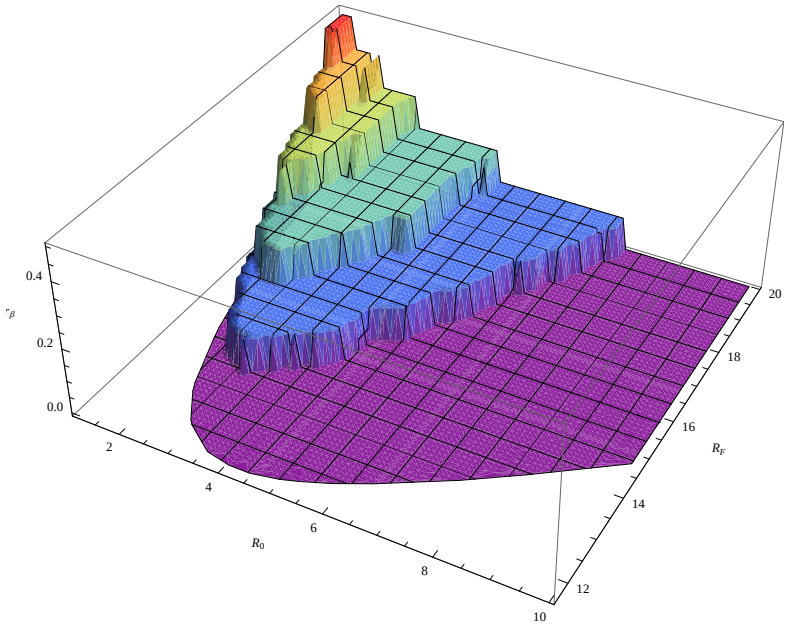

Figure 5.4: Phase space of parameters $R_0 \times R_F \times r_\beta$ in which we get two peaks for $\alpha = 0.05$. In the simulations we explored the region $1 < R_0 \leq 10$, $0.1 < R_F < 20$ and $0 \leq r_\beta < 1$

The new interaction, although intuitively simple, significantly complicates the dynamics of the model. In particular, for some parameter values, we observed two epidemic peaks as shown in Figure (5.3). This non-trivial behavior can be easily understood. Fear self reinforces until it severely depletes the reservoir of susceptibles, causing a decline in new cases. As a result, people are lured in to a false sense of security and return to their normal behavior (recover from fear) causing a second epidemic peak that can even be larger than the first one. Some authors believe that a similar process occurred during 1918 pandemic resulting in multiple epidemic peaks (114).

A better understanding of the conditions in which this is possible is of obvious practical importance. In Figure (5.4) we show the, numerically found, region of parameter space that results in two epidemic peaks for a fixed value of α. It is clear how this region is reduced and shifted toward regions of larger R_F and smaller R_0 as r_β increases. This effect is due to the reduction on the protection that scared individuals experiece as r_β increases. In the limit $r_\beta \equiv 1$ the model is indistinguishable from an SIR model that does not allows a second peak.

After the end of the epidemic, the system enters in the disease-free equilibrium. It is easy to prove that:

$$S^F_{I=0}(t) = \frac{\gamma}{R_F - 1 + \Theta e^{-\gamma \mu_F t}}, \tag{5.16}$$

where we defined

$$\gamma = R_F(1 - R_\infty) - 1, \tag{5.17}$$

and Θ is a constant function of the parameters of the model. Interestingly, there are two possible disease-free equilibria :

$$\gamma \leq 0 \rightarrow (S_\infty, S^F_\infty, I_\infty, R_\infty) = (1 - R_\infty, 0, 0, R_\infty), \tag{5.18}$$

where fear dies along with the disease, and:

$$\gamma > 0 \rightarrow (S_\infty, S^F_\infty, I_\infty, R_\infty) = (\frac{R_\infty}{R_F - 1}, 1 - \frac{R_F R_\infty}{R_F - 1}, 0, R_\infty), \tag{5.19}$$

where fear and changes of behavior persist even after the end of the epidemic. $R_F > 1$ is a necessary but not sufficient condition to have an endemic state of fear, while $R_F \leq 1$ is sufficient to avoid an endemic state of fear. Unfortunately, the parameter γ is an implicit function of the whole dynamics through the epidemic size R_∞. The presence of and endemic state, a societal memory, of fear is a quite interesting feature of the model. It indicates that an event localized in time is capable of permanently modifying society with interesting consequences. In the case of a second epidemic, the presence of part of the population already in the compartment S^F reduces the value of R_0:

$$R_0 = \frac{\beta}{\mu}\left[r_\beta + \frac{R_1(1 - r_\beta R_F)}{R_F - 1}\right], \tag{5.20}$$

and consequently the severity of second epidemic. We can write

$$R_0 < R_0^{SIR} = (1 - R_1)\frac{\beta}{\mu}, \tag{5.21}$$

where R_1 is the population already immune from the first pandemic. To prove the last inequality we have to show that

$$r_\beta + \frac{R_1(1 - r_\beta R_F)}{R_F - 1} < 1 - R_1, \tag{5.22}$$

that means

$$\frac{R_1(1 - r_\beta R_F)}{R_F - 1} < 1 - R_1 - r_\beta. \tag{5.23}$$

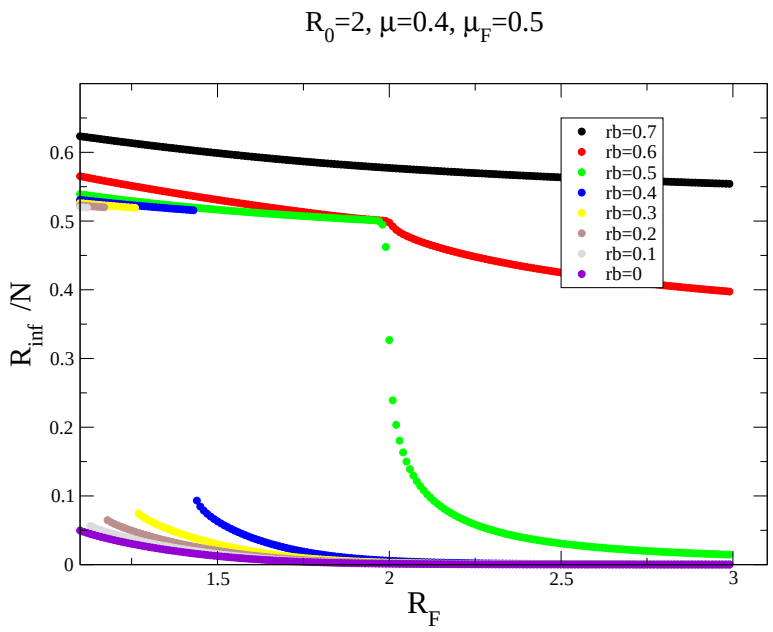

Figure 5.5: Reduction of the epidemic size as a function of R_F for different values of r_β, $R_0 = 2$, $\mu = 0.4$, $\mu_F = 0.5$ and $\alpha = 0.05$

The functions in the first and second member are monotonic functions of r_β and of course for $r_\beta = 1$ they have the same value $-R_1$. To prove our preposition we have just to confront the slope of the function an show that

$$\frac{R_1 R_F}{R_F - 1} < 1, \tag{5.24}$$

that means

$$R_1 < 1 - \frac{1}{R_F}. \tag{5.25}$$

The last inequality is always true, provided our assumption $\gamma > 0$. This is an important result. It is clear how an endemic state of fear in the population reduces the impact of a second outbreak.

Epidemic reduction and phase transition

Even in this model, the presence of fear results in a reduction on the epidemic size as showed in Figure (5.5). In this case the self-reinforcing mechanism create a more

complicate phase space that allows first order phase transitions for a wide range of r_β. In particular this range is characterized by small values ($0 \leq r_\beta \leq 0.5$ in the case shown in the Figure), when the protection of feared people is relatively large. The transition to fear due to infected people is progressively suppressed and the SS^F interaction becomes increasingly important until the critical point in which the fear spreads fast enough to eradicate the disease, due to the depletion of susceptibles. Intriguingly, as shown in Figure (5.6), the critical value of R_F corresponds to the first value characterized by $\gamma > 0$. The phase transition is clearly related to the presence of an endemic state of fear. The extremely rich phase space of the model can be divided in two totally different regions. For set of parameter below the critical point no endemic state of fear is allowed and there is a mild reduction on the epidemic size. Above the critical point we have a huge drop on the epidemic size and an endemic state of fear. In this region of the phase space the transition into fear is the leading order and the reduction provided by the state of awareness is enough to sustain a permanent state of fear. This is important for two reasons: we have a strong reduction in the cumulative number of infected individuals and in the case of a new epidemic the memory of the system reduces the spreading of the disease shifting the reproductive number towards smallest values. These are very interesting properties of the model due to the self reinforcing mechanism that clearly create non trivial behaviors in the dynamics. We have tried different analytical approaches to get more insight into the phase transition. This is still an open issue that will be matter of next intensive studies.

5.3 Mass-media effect

The final fear inducing process we considered is the diffusion of awareness through mass-media. To increase ratings, mass-media widely advertises the progress of the epidemic making even people that have never contacted a diseased or a scared person be aware of the disease. In this formulation, the rate of the transition to fear can be thought to be related to the absolute value of infectious people, instead of the fraction of total population. The coupling between S and I, can be written as:

$$\beta_F S(t)\frac{I(t)}{N} \rightarrow \beta_F S(1 - e^{-\alpha I(t)}), \tag{5.26}$$

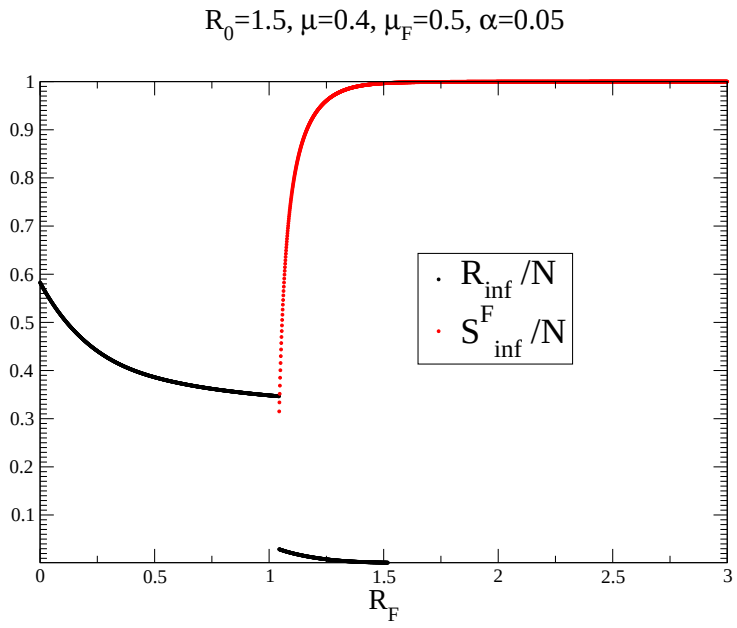

Figure 5.6: S^F/N and R/N for $r_\beta = 0$, $R_0 = 1.5$, $\mu = 0.4$, $\mu_F = 0.5$ and $\alpha = 0.05$

where $0 < \alpha \leq 1$. For small values of α we have the *pseduo mass action* interaction of the first order in α:

$$\beta_F S(1 - e^{-\alpha I(t)}) = \beta_F S(t) \left[\alpha I(t) + \mathcal{O}(\alpha^2)\right], \tag{5.27}$$

and the equations of the model become:

$$\begin{aligned}
d_t S(t) &= -\beta S(t)\frac{I(t)}{N} - \beta_F S(t)\left[1 - e^{-\alpha I(t)}\right] + \mu_F S^F(t)\left[\frac{S(t)+R(t)}{N}\right], \\
d_t S^F(t) &= -r_\beta \beta S^F(t)\frac{I(t)}{N} + \beta_F S(t)\left[1 + e^{\alpha I(t)}\right] - \mu_F S^F(t)\left[\frac{S(t)+R(t)}{N}\right], \\
d_t I(t) &= -\mu I(t) + \beta S(t)\frac{I(t)}{N} + r_\beta \beta S^F(t)\frac{I(t)}{N}, \\
d_t R(t) &= \mu I(t).
\end{aligned}$$

Assuming that the epidemic spreads faster than epidemic awareness the reproductive number R_0 is the same as in previous models: $R_0 = \beta/\mu$. In the limit $R_F \to \infty$ it is easy to understand that $R_0 \to R_0 r_\beta$. Awareness, magnified by mass media, spreads instantaneously and all susceptibles immediately move to the S^F compartment. The

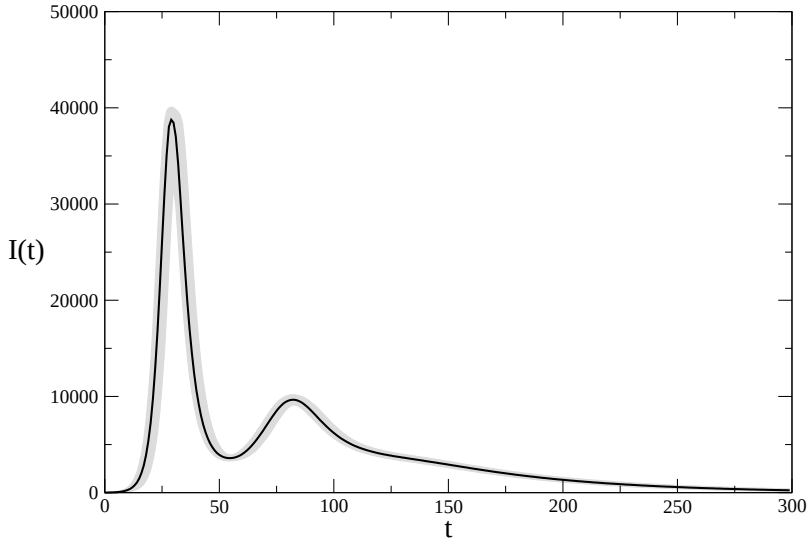

Figure 5.7: Median and 95 % reference range of $I(t)$ for $r_\beta = 0.1$, $R_0 = 2$, $\mu = 0.4$, $\mu_F = 0.1$, $\alpha = 10^{-5}$ and $R_F = 4.2$

early time evolution of S^F, is now much more complex:

$$S^F(t) \sim \beta_F e^{-\mu_F t} \int \left(1 - e^{-\alpha I_0 e^{\mu(R_0-1)t}}\right) e^{\mu_F t} dt. \tag{5.28}$$

This expression is integrable in two regimes, small α or if $\mu(R_0 - 1) = \mu_F$. In the first case:

$$S^F(t) \sim \frac{\beta_F \alpha I_0}{\mu(R_0 - 1) + \mu_F} \left[e^{\mu(R_0-1)t} - e^{-\mu_F t}\right]. \tag{5.29}$$

As in the first model, if $\mu(R_0 - 1) > \mu_F$ the awareness spreads or dies off along side the epidemic. In the second case the early time behavior reads:

$$S^F(t) \sim \frac{\beta_F}{\mu_F}\left(1 - e^{-\mu_F t}\right) + \frac{\beta_F I_0}{\alpha \mu_F} e^{-\mu_F t}\left(e^{-\alpha e^{\mu_F t}} - e^{-\alpha}\right). \tag{5.30}$$

Interestingly, even the phase space of this model is much more richer than the first one proposed. We can obtain two peaks in the I profiles, as shown in Figure (5.7), but the disease free equilibrium does not allows an endemic state of fear. The transition to fear is based just on the presence of infected individuals. As soon the epidemic dies out the

in-flow to the S^F compartment stops, while the out-flow continues to allow people to recovered from fear. When the media coverage vanishes, so does the fear it spreads.
Even in this model the effect of the fear brings a reduction of the epidemic size. The reduction is function of α and of all the parameters. As α increase the transitions into fear becomes faster. Fear people are more protect from the disease and the epidemic size decrease. Keeping fix α and increasing R_F the epidemic size is reduced as well, at least before the stationary state. The asymptotic value of R_∞ as a function of R_F depends on the product $r_\beta\beta/\mu$. If this product is bigger than 1, Figure (5.8), the asymptotic value is the epidemic size of an SIR model with $\beta' = \beta r_\beta$. If instead the product is smaller the 1, Figure (5.9), the asymptotic value is zero: the scale time of the spreading of awareness is infinitively faster than the disease one. This process can be thought as an SIR with an reproductive number smaller that 1.

The modeling of social distancing during an epidemic outbreak is still a big issue in epidemiology. In this Chapter we introduced a general framework with different mechanism considering the spread of awareness of a disease as another contagion process. Three mechanism were proposed:

- in the first, basic, model the social disruption is just related to the fraction of infected individuals in the population. As a contagion process people enter in the compartment of feared people interacting with infected individual with a rate β_F. While in this compartment they reduce their contacts, gaining a reduction on the probability to get sick. People can recover from this state of awareness meeting susceptible or recovered individuals with rate μ_F. For a wide range of parameters the epidemic size is reduced. This effect is merely due to the reduction of the contacts.

- In the second model we added the possibility that susceptible people get into the S^F compartment after an interaction with people already feared. This apparently simple interaction allows a self-reinforcement of fear. We discovered that the phase space of this model is much more rich than the first one. We found a range of parameters with two peaks in the incidence curve and others in which is present a disease free equilibrium with an endemic state of fear, related to a first order phase transition. These features are extremely interesting and show a non trivial

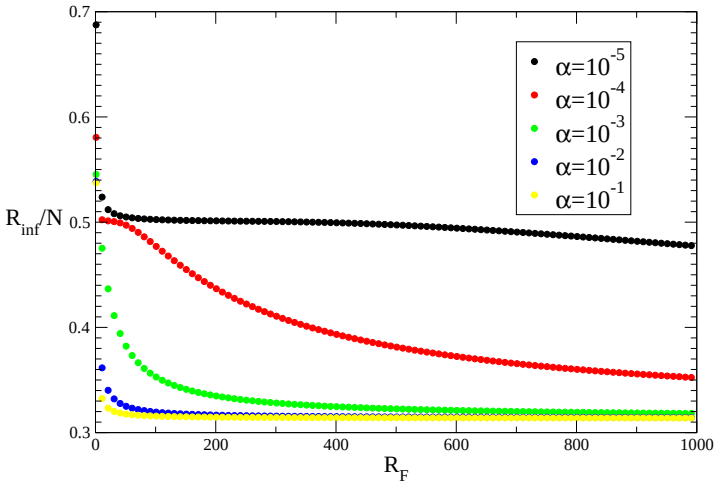

Figure 5.8: Reduction of the epidemic size as a function of R_F for different values of α, $R_0 = 2$, $\mu = 0.4$, $\mu_F = 0.5$ and $r_\beta = 0.6$

phase space. The presence of endemic state or memory in the system is important characteristic that in the case of a second epidemic spreading reduces the value of the reproductive number. At this stage the study of these properties has been just phenomenological. The identification of a most clear region in which these phenomena are present is an open issue, material for future work.

- In the last model we introduced a mass-media spreading effect. We modeled the spread of awareness considering just the absolute number of infected individuals. An exponential coupling was proposed. We found that even in this model the phase space is much more rich than in the first one. A region of parameters with two peaks in the incidence curve is present as well as a reduction on the epidemic size. Instead in this case due to the absence of any self-reinforcing term an endemic state of fear is not allowed.

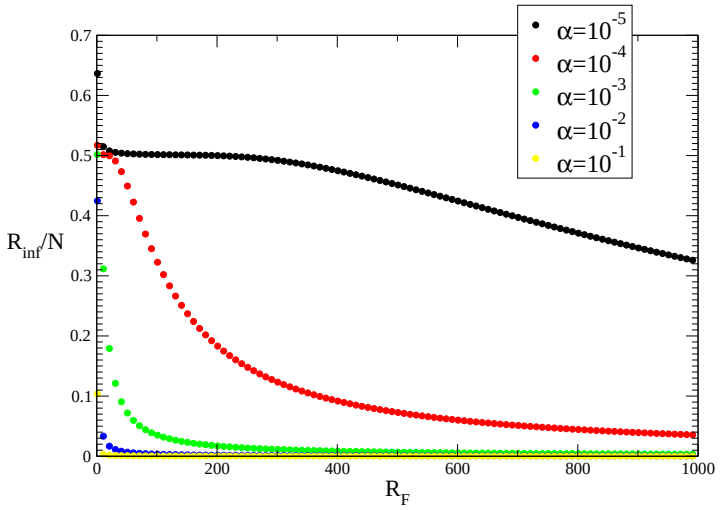

Figure 5.9: Reduction of the epidemic size as a function of R_F for different values of α, $R_0 = 2$, $\mu = 0.4$, $\mu_F = 0.5$ and $r_\beta = 0.4$

6

Metapopulation Models

In science one tries to tell people, in such a way
as to be understood by everyone, something
that no one ever knew before.
But in the case of poetry,
it's the exact opposite.
P. Dirac

Contents

In the Chapters 2 and 5 we studied systems in which each node of the network correspond to a single individual or a single population with a homogeneous mixing approximation. Recently the effect of heterogeneous connectivity patterns has been studied in the case in which each node of the system may be occupied by any number of particles and the links allow for the displacement of particles form one node to the other. In a epidemic framework, particles represent people moving between different locations, such cities or urban areas and the reaction processes between individuals present in the same location are governed by infection dynamics. These models are called metapopulations epidemic models and are based on the detailed knowledge of the spatial structure of the environment and of transportation infrastructures, movement

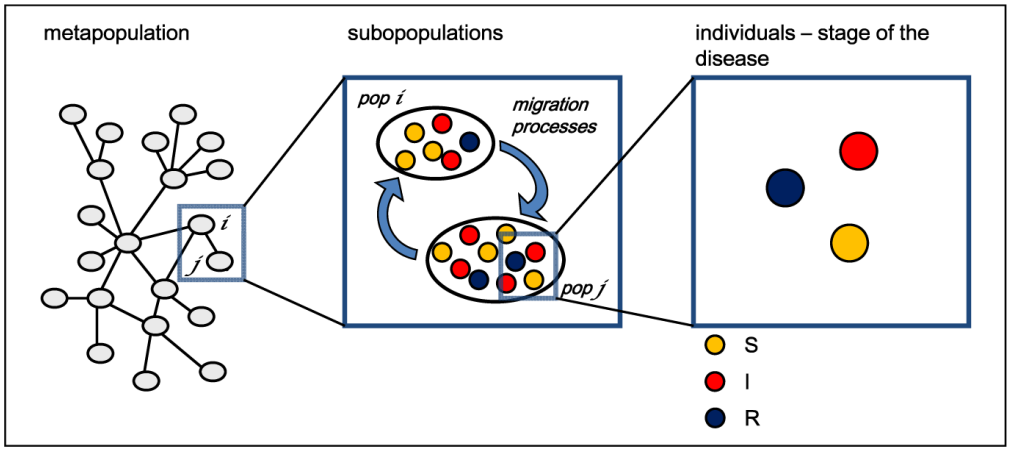

Figure 6.1: Representation of a metapopulation model. Each node of the system contains a population of individuals who are characterized with respect to their stage of the disease. In this case we are considering Susceptible, Infected and Recovered indicated in different color in the picture. Individuals can diffuse from a subpopulation/node to another on the network of connections among subpopulations. Figure courtesy of prof. A. Vespignani.

patterns and traffic networks (18; 19; 20; 21; 122; 123; 124).

Individuals within each subpopulation are divided into classes denoting their state with respect to the modeled disease (76). These subpopulations are connected with a certain topology, are coupled, and individuals in the same location may get into contact changing their state according to the infection dynamics. The coupling among subpopulations is the result of the movement of individuals from one subpopulation to the other. A sketch of the metapopulation approach is shown in Figure (6.1). Each node i is connected to other k_i nodes according to its degree resulting in a network with degree distribution $P(k)$ and distribution moments $\langle k^\alpha \rangle = \sum_k k^\alpha P(k)$.

A key point is to evaluate the force of infection generated by the infectious individuals in subpopulation j on the individuals in subpopulation i (19; 20; 125; 126; 127; 128). Realistic descriptions are provided by explicit mechanistic approaches, in which detailed rate of traveling/commuting obtained from data, or from empirical fit to gravity law models, are included (21; 129).

A typical assumption is to set the diffusion process as Markovian. As we described in Chapter 2 this implies that the movements of individuals have no memory. Individuals are not labeled according to their original subpopulation, so they move without having

memory of their origin. At each time step the movement of individuals is given according to a matrix d_{ij} that give us the probability that an individual in the subpopulation i will travel to the subpopulation j. Let us define w_{ij} the traffic among subpopulations. We define:

$$d_{ij} \sim \frac{w_{ij}}{N_i}. \tag{6.1}$$

These probabilities in realistic models are obtained from real data (16; 21; 130; 131; 132; 133; 134; 135; 136).

Let us consider again a representation of the quantities using a degree block approach. So the average number of individuals in node of degree k will be:

$$N_k = \frac{1}{V_k} \sum_{i|k_i=k} N_i, \tag{6.2}$$

where V_k is the number of nodes with degree k. Let us consider the diffusion rate between a subpopulation of degree k and k' as $d_{kk'}$. The rate at which individuals leave a subpopulation with degree k is:

$$p_k = \sum_{k'} P(k'|k) d_{kk'}. \tag{6.3}$$

Using a typical mean-field dynamical equation we can write the variation in time of the subpopulations in each degree block:

$$\partial_t N_k(t) = -p_k N_k(t) + k \sum_{k'} P(k'|k) d_{k'k} N_{k'}(t), \tag{6.4}$$

where the first term take into account that a fraction p_k of individuals moves out of the node, and the second term is proportional to the degree k times the average number of particles coming from each neighbors. As usual this term depends on the rate of diffusion between degree's classes $d_{kk'}$, on the number of individuals present in each neighbors subpopulations and, on the conditional probability to find a link among degree's classes. Assuming uncorrelated networks we have:

$$\partial_t N_k(t) = -p_k N_k(t) + \frac{k}{\langle k \rangle} \sum_{k'} k' P(k') d_{k'k} N_{k'}(t). \tag{6.5}$$

We can solve this set of differential equation defining the type of diffusion processes. As we will discuss in details in Chapter 7 a relevant networks in the field of epidemic is given by the airline transportation network. In this system for each direct flight

connection between airports i and j a weight w_{ij} is assigned, which corresponds to the number of available seats or passengers in that route. The traffic shows a probability distribution $P(w)$ varying over six orders of magnitude, so a heavy-tail behavior. As shown in Ref. (33) is possible to describe the average of weight along the connections between subpopulations with degree k and k' as function of their degree:

$$< w_{kk'} >= w_0(kk')^\theta, \tag{6.6}$$

where w_0 and θ depend on the specific system. For the world-wide air transportation network we have $\theta = 0.5$. An important quantity related to this is the total average traffic per unit time of the subpopulation with degree k:

$$T_k = \sum_{k'} w_0(kk')^\theta = Ak^{(1+\theta)}, \tag{6.7}$$

where A depends on the system we are considering. A realistic process can considers the movement of individuals to be proportional to the traffic intensity along a given edge. In this case we can consider a heterogeneous diffusion rate:

$$d_{kk'} = p\frac{w_0(kk')^\theta}{T_k}, \tag{6.8}$$

where we set $p_k = p \ \forall k$. The diffusion rate is constant in each subpopulation but the individuals move on each connection proportionally to the the traffic on the connection. Using this into the (6.5) we get:

$$\partial_t N_k(t) = -pN_k(t) + pk^{(1+\theta)}\frac{w_0}{A\langle k\rangle}\sum_{k'} P(k')N_{k'}(t). \tag{6.9}$$

The stationary solution $\partial_t N_k(t) = 0$ does not depend upon the diffusion rate p, that just set the time scale at which the equilibrium is reached, has the solution:

$$N_k = k^{(1+\theta)}\frac{w_0}{A\langle k\rangle}\bar{N}, \tag{6.10}$$

where $\bar{N} = \sum_k P(k)N_k(t)$ represent the average subpopulations size. Using $A =< k^{1+\theta} > w_0/ < k >$ for uncorrelated networks we get:

$$N_k = \frac{k^{(1+\theta)}}{\langle k^{(1+\theta)}\rangle}\bar{N}. \tag{6.11}$$

In the stationary limit the population of each node scales with the node degree. If we fix $\theta = 0$ we recover the homogeneous diffusion case in which $d_{kk'} = d_k = p/k$ and:

$$N_k = \frac{k}{\langle k \rangle} \bar{N}, \tag{6.12}$$

the subpopulation is this case is fixed from the topological fluctuations. It is clear now the role of θ that takes into account traffic fluctuations.

We can now relax the condition that we used in the previous case $p_k = p \ \forall k$. We can consider that the diffusion rate is general inversely proportional to the population size and proportional to the total flow among the node. An important empirical evidence shows that in large-scale transportation we have:

$$\partial_t N_i = \sum_j (w_{ji} - w_{ij}) = 0, \tag{6.13}$$

so the matrix of weights can be considered as symmetric. Fixing the diffusion rate as $p_k = T_k/N_k$ we have

$$d_{kk'} = \frac{w_0 (kk')^\theta}{N_k}, \tag{6.14}$$

then:

$$\partial_t N_k(t) = -T_k + k^{(1+\theta)} w_0 \frac{\langle k^{1+\theta} \rangle}{\langle k \rangle}. \tag{6.15}$$

For uncorrelated networks we know that $T_k = k^{(1+\theta)} w_0 \langle k^{1+\theta} \rangle / \langle k \rangle$ so we recover by definition $\partial_t N_k(t) = 0$. This is an important result. In the previous subsection each individual has the same diffusion rate p and as effect of that the subpopulation reached a fix size after a transient. In this case instead a population dependent diffusion process does not fix the subpopulation size which can be given as a free parameter of the model.

6.1 Epidemic spreading and the invasion threshold

Our goal is to explore the epidemic behavior in metapopulations models. To start we need to explicitly consider a disease dynamics inside each subpopulation. We will consider a standard compartmentalization described in Chapter 2. A crucial quantity for the spreading in the single population is the reproductive number R_0. Only if $R_0 > 1$ any epidemic will spread across a non-zero fraction of the population. At the metapopulation level, the epidemic behavior on the global scale, is related also to the diffusion process of individuals. The effects due to finite size of subpopulations and the stochastic nature

of the diffusion might have a crucial role. It is important then to consider the discrete nature of individuals. Each subpopulation may or may not transmit the infection to another subpopulation it is in contact with. This depends on the occurrence that at least one infected individual will travel to a non-infected subpopulation. The spreading process across subpopulations is then related to the diffusion rate of individuals and the total number of individuals who will experience the infection (137; 138). For $R_0 < 1$ the disease has not hope to spread among the subpopulations. For $R_0 > 1$ the disease may or not spread into the subpopulations. For a SIS model, in which the number of infected individuals reached a stationary state, the epidemic will eventually spread to different populations, since locally endemic. In a SIR model instead, the global spreading is strongly related not just on the local outbreak but even on the diffusion rate that has to be big enough to allow infected people to travel into different subpopulations before recover. In this case the simply reproductive number is not enough to describe the global threshold. A new predictor is introduced R_*. It is called *global invasion threshold*. In the next sections we will evaluate this quantity for homogeneous and heterogeneous metapopulation networks.

As a general framework let us consider a metapopulation system in which a seed (infected individual) is introduced in a subpopulation of degree k and size N_k. Let us consider the case in which $R_0 > 1$. As we said in Chapter 2 there is probability of extinction of the epidemic equal to:

$$P^{ext} = \frac{1}{R_0}. \tag{6.16}$$

In general the epidemic will affect a finite fraction of the population with non-zero probability. In case of a global outbreak we can consider the number of infected people during the evolution of the epidemic equal to αN_k, where α is function of the specific disease model and parameters used. Considering a SIR model each infected individual will stays in the infectious state on average a time μ^{-1}. During this time it can travel into the neighboring subpopulation of degree k' with rate $d_{kk'}$. As a first approximation we can evaluate the average number on new seeds coming from a subpopulations k into a connected subpopulations k' using:

$$\lambda_{kk'} = d_{kk'} \frac{\alpha N_k}{\mu}. \tag{6.17}$$

Let us define D_k^0 as the number of diseased subpopulation of degree k at generation 0. Those are experiencing an outbreak at the beginning of the spreading. Each of those

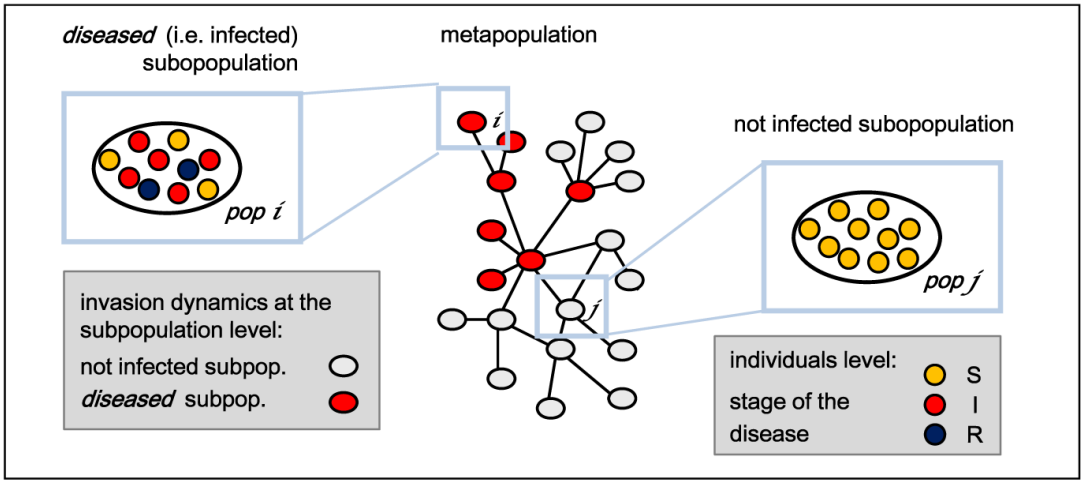

Figure 6.2: Representation of the invasion dynamics at the level of the subpopulations. In a metapopulation approach the system is considered in a coarse grained perspective as a network where each node represents a subpopulation which can be infected (*diseased*) if it is reached by the virus carried by infected individuals diffusing on the system. Figure courtesy of prof. A. Vespignani.

subpopulation during the course of the epidemics can infected another one defining the set D_k^1 and so on. This process is shown in Figure (6.2). We have a branching process where the n^{th} generation of infected subpopulations of degree k will be denoted as D_k^n (137; 139; 140). In the early time we can assume, with good approximation, that the number of subpopulations already affected by an outbreak is small. We can study then the evolution of the diseased subpopulations using a tree-like approximation relating D_k^n with just D_k^{n-1}.

6.1.1 Global invasion threshold in homogeneous metapopulation networks

Let us consider a particular case in which the metapopulation system is a homogeneous random graph. Each subpopulation has the same degree $k = \bar{k}$ and population $\bar{N}$. Dropping the subscript index k we have:

$$D^n = D^{n-1}(\bar{k} - 1)\left[1 - \left(\frac{1}{R_0}\right)^{\lambda_{\bar{k}\bar{k}}}\right]\left(1 - \frac{D^{n-1}}{V}\right). \tag{6.18}$$

In this equation we are assuming that each infected subpopulation of the $(n-1)^{th}$ generation will seed with infected individuals a number of other subpopulations depending on the number of neighbors (minus the one which originally transmitted the disease) $\bar{k}-1$, times the probability that the subpopulation is not already diseased $(1-D^{n-1})/V$, and the probability that the new seeded subpopulation will experience an outbreak $\left(1-R_0^{-\lambda_{\bar{k}\bar{k}}}\right)$ (80). We can consider now the simplest case of homogeneous diffusion of individuals $d_{\bar{k}} = p/\bar{k}$ that yields:

$$\lambda_{\bar{k}\bar{k}} = p\bar{N}\frac{\alpha}{\mu}\frac{1}{\bar{k}}. \tag{6.19}$$

In order to get an explicit result we will consider $R_0 \to 1$ so that:

$$\left[1-\left(\frac{1}{R_0}\right)^{\lambda_{\bar{k}\bar{k}}}\right] \simeq \lambda_{\bar{k}\bar{k}}(R_0-1). \tag{6.20}$$

$D^{n-1}/V \ll 1$ is a further possible approximation. Using this:

$$D^n = p\bar{N}\alpha\mu^{-1}\frac{\bar{k}-1}{\bar{k}}(R_0-1)D^{n-1}. \tag{6.21}$$

We will have an outbreak only if the quantity:

$$R_* = p\bar{N}\alpha\mu^{-1}\frac{\bar{k}-1}{\bar{k}}(R_0-1) > 1. \tag{6.22}$$

This is the *global invasion threshold*. It is possible to get the diffusion rate of individuals for the global spread of the epidemic in the metapopulation system:

$$p\bar{N} \geq \frac{\bar{k}}{\bar{k}-1}\frac{\mu}{\alpha}(R_0-1)^{-1}. \tag{6.23}$$

This threshold tell us that there is a minimum diffusion rate to ensure that on average each subpopulation can seed more that one neighboring subpopulation. Following Ref. (14) is possible to approximate the constant α for an SIR model in the case of R_0 close to 1:

$$\alpha \simeq 2\frac{\mu}{\beta}(1-\frac{\mu}{\beta}) = \frac{2(R_0-1)}{R_0^2}. \tag{6.24}$$

Using this in the previous equation we can write:

$$p\bar{N} \geq \frac{\bar{k}}{\bar{k}-1}\frac{\mu R_0^2}{2(R_0-1)^2}. \tag{6.25}$$

It is easy to see that for R_0 very close to the unity in the single population, the diffusion rate needed to get a global outbreak has to be very large. It is important to stress that all these expression for R_* are valid in the limit $R_0 \sim 1$. The expansion presented are no longer valid for larger value of R_0. In that case they can be obtained only in the form of complicated implicit expressions.

6.1.2 Global invasion threshold in heterogeneous metapopulation networks

In the case of heterogeneous metapopulation networks we have to consider explicitly the degree and population heterogeneities. A population of degree k' can seed $k'-1$ subpopulations, we have to consider the conditional probability that a subpopulation of degree k' will be connected to a subpopulation of degree k. Using the previous approximations $R_0 - 1 \ll 1$ and $(1 - D_k^{n-1}/V_k) \simeq 1$ we get:

$$D_k^n = \sum_{k'} D_{k'}^{n-1}(k'-1)\lambda_{k'k}(R_0-1)P(k|k')\left(1 - \frac{D_k^{n-1}}{V_k}\right). \tag{6.26}$$

We can simplify the expression assuming that degree correlations can be neglected:

$$D_k^n = \frac{kP(k)}{\langle k\rangle}(R_0-1)\sum_{k'} D_{k'}^{n-1}(k'-1)\lambda_{k'k}. \tag{6.27}$$

At this point the next step is to specify the form of $\lambda_{k'k}$. Let us start with a heterogeneous diffusion rate (6.8) and the stationary value for each subpopulation (6.11). We have:

$$\lambda_{k'k} = \frac{p\langle k\rangle}{\langle k^{(1+\theta)}\rangle}\frac{\alpha}{\mu}\frac{(k)^\theta}{k'}N_{k'} = \frac{p\langle k\rangle}{\langle k^{(1+\theta)}\rangle^2}\frac{\alpha}{\mu}(kk')^\theta\bar{N}. \tag{6.28}$$

Using this in the (6.26) we get:

$$D_k^n = (R_0-1)\frac{k^{1+\theta}P(k)}{\langle k^{1+\theta}\rangle^2}\frac{p\bar{N}\alpha}{\mu}\sum_{k'} D_{k'}^{n-1}k'^\theta(k'-1). \tag{6.29}$$

We can define $\Theta^n = \sum_{k'} D_{k'}^n k'^\theta(k'-1)$ and obtain a iterative form:

$$\Theta^n = (R_0-1)\frac{\langle k^{2+2\theta}\rangle - \langle k^{1+2\theta}\rangle}{\langle k^{1+\theta}\rangle^2}\frac{p\bar{N}\alpha}{\mu}\Theta^{n-1}. \tag{6.30}$$

We will have an increase of infected diseased subpopulation if:

$$R_* = (R_0-1)\frac{\langle k^{2+2\theta}\rangle - \langle k^{1+2\theta}\rangle}{\langle k^{1+\theta}\rangle^2}\frac{p\bar{N}\alpha}{\mu} > 1. \tag{6.31}$$

It is interesting to notice that this expression differs from the homogeneous one for a factor related to the topology of the network. For heavy-tailed distribution of degree this term is vanishing in the limit of infinite size. The heterogeneity even in this case is favoring the global spread by lowering the global invasion threshold as we saw in the Chapter 2 in the case of contact pattern heterogeneity.
Let us consider now a more realistic case in which the diffusion rate is set to be proportional to the ratio between traveling people and population size:

$$p_k = \frac{T_k}{N_k}. \tag{6.32}$$

We have shown that in this case we have stationary populations sizes independent on the diffusion process. We can set:

$$\lambda_{kk'} = w_0 (kk')^\theta \alpha \mu^{-1}. \tag{6.33}$$

Using the approximations used in the previous cases we get:

$$D_k^n = (R_0 - 1) \frac{k^{1+\theta} P(k)}{\langle k \rangle} \frac{w_0 \alpha}{\mu} \sum_{k'} D_{k'}^{n-1} k'^\theta (k' - 1). \tag{6.34}$$

Considering as before the auxiliary function θ we can write:

$$\Theta^n = (R_0 - 1) \frac{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle}{\langle k \rangle} \frac{w_0 \alpha}{\mu} \Theta^{n-1}, \tag{6.35}$$

from which we obtain the global invasion condition:

$$R_* = (R_0 - 1) \frac{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle}{\langle k \rangle} \frac{w_0 \alpha}{\mu} > 1. \tag{6.36}$$

The mobility threshold reads as:

$$w_0 \geq \frac{\langle k \rangle}{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle} \frac{\mu R_0^2}{2(R_0 - 1)^2}, \tag{6.37}$$

also in this case the mobility threshold is extremely lowered by the topological fluctuations of the network. As shown in Ref. (33) using real data and this last expression an epidemic carried by air travelers would reach global portion if the average number of travelers per day would be larger than:

$$w_0^{real} \sim 3 \times 10^{-3}. \tag{6.38}$$

The average traffic on a given connection has a minimum corresponding to $\sim 10^{-2}$. Our air transportation networks is almost one order of magnitude above threshold, and how we experienced recently during the pandemic of influenza H1N1 and will discuss in details in the Chapter 7, really vulnerable.

6.1.3 Epidemic behavior above the invasion threshold

In this section will imagine a system above the invasion threshold. With a finite probability the epidemics will affect a macroscopic fraction of subpopulations. In this limit we can study the the epidemic spreading using deterministic equations. Following Ref. (23) we will describe the disease dynamics of an SIR model in each subpopulation as a reaction process. As we did in Chapter 2 we will use a mass-action transmission rate. The probability that a susceptible has a contact with an infectious individual is equal to the density of infected within the subpopulation:

$$\lambda_j = \beta\Gamma_j = \beta\frac{I_j S_j}{N_j}, \tag{6.39}$$

where Γ_j is usually called interaction kernel. Instead to consider just the number diseased subpopulations here we consider the degree block variables for I and S:

$$I_k = \frac{1}{V_k}\sum I_j; \qquad S_k = \frac{1}{V_k}\sum S_j. \tag{6.40}$$

We can write analogously the reaction kernel for each degree block as:

$$\Gamma_k = S_k\frac{I_k}{N_k}. \tag{6.41}$$

At the end of the reaction-diffusion process between t and $t+\Delta t$ we have the following master equation:

$$\partial_t I_k = -p_k I_k + (1-p_k)\left[-\mu I_k + \beta\Gamma_k\right] + k\sum_{k'} P(k'|k)d_{k'k}\left[(1-\mu)I_{k'} + \beta\Gamma_{k'}\right]. \tag{6.42}$$

The negative terms on the right side are due to the diffusion in other subpopulations ($p_k I_k$) and to the infected individuals that do not diffuse away but recovered ($(1-p_k)\mu I_k$). The positive term are due to susceptible individuals that do not diffuse but in contact with infected get sick ($(1-p_k)\beta\Gamma_k$) plus a term that take into account the infected individuals that from the neighbor come in but do not recovered and the new generated one. Considering the uncorrelated case we have:

$$\partial_t I_k = -p_k I_k + (1-p_k)\left[-\mu I_k + \beta\Gamma_k\right] + \frac{k}{\langle k\rangle}\sum_{k'} k'P(k')d_{k'k}\left[(1-\mu)I_{k'} + \beta\Gamma_{k'}\right]. \tag{6.43}$$

At the early stage of the epidemics, as we discussed in Chapter 2, we can neglect all the term order of $\mathcal{O}(I^2)$ so that the kernel transition can be written as:

$$\Gamma_k = \frac{(N_k - I_k - R_k)I_k}{N_k} \simeq I_k. \tag{6.44}$$

Using this in the (6.43) we have:

$$\partial_t I_k = -p_k I_k + (1-p_k)(\beta-\mu)I_k + \frac{k}{\langle k\rangle}\sum_{k'} k' P(k') d_{k'k}\left[(1-\mu+\beta)I_{k'}\right]. \tag{6.45}$$

In order to get explicit solutions we have to consider specific diffusion processes already introduced in the previous sections.

Let us first consider the case of traffic dependent mobility rates and a uniform p as in the (6.8). By considering $\bar{I} = \sum_{k'} P(k') I_{k'}$, it is easy to show:

$$\partial_t I_k = -pI_k + (1-p)(\beta-\mu)I_k + p\frac{k^{(1+\theta)}}{\langle k^{(1+\theta)}\rangle}\left[(1-\mu+\beta)\bar{I}\right], \tag{6.46}$$

averaging both terms over $P(k)$ we obtain:

$$\partial_t \sum_k P(k) I_k = \partial_t \bar{I} = (\beta-\mu)\bar{I}, \tag{6.47}$$

that has the simple solution:

$$\bar{I} = \bar{I}(0) e^{(\beta-\mu)t}. \tag{6.48}$$

$\bar{I}(0)$ is the initial number of infected individuals in the metapopulation. We have an increase of the number of infected individuals if $\beta > \mu$. We recover the epidemic threshold for the SIR model $R_0 = \beta/\mu > 1$. At the deterministic level the global threshold is equivalent to the condition for each single population. It is possible to solve the early time for all subpopulations for a given degree block using the solution for $\bar{I}$ in the (6.46):

$$I_k(t) = A\frac{k^{1+\theta}}{\langle k^{1+\theta}\rangle} e^{(\beta-\mu)t} + C_k e^{[(1-p)(\beta-\mu)-p]t}, \tag{6.49}$$

where A and C_k are fixed by the initial conditions. If the seeds are distributed only in the k_0-block so that $I_k(0) = \delta_{k,k_0}\bar{I}(0)/P(k_0)$ we have:

$$A = \bar{I}(0) \quad \text{and} \quad C_k = \bar{I}(0)\left(\frac{\delta_{k,k0}}{P(k_0)} - \frac{k^{1+\theta}}{\langle k^{1+\theta}\rangle}\right). \tag{6.50}$$

If the seeds are homogeneously distributed instead:

$$A = \bar{I}(0) \quad \text{and} \quad C_k = \bar{I}(0)\left(1 - \frac{k^{1+\theta}}{\langle k^{1+\theta}\rangle}\right). \tag{6.51}$$

We can now consider the case in which a population dependent diffusion rate is used: $p_k = T_k/N_k$. Considering $T_k = k^{1+\theta} w_0 \langle k^{1+\theta}\rangle / \langle k\rangle$ we obtain:

$$\partial_t I_k = -p_k I_k + (1-p_k)(\beta-\mu)I_k + \frac{k^{1+\theta}}{\langle k^{1+\theta}\rangle}(1+\beta-\mu)\Omega, \tag{6.52}$$

where we defined $\Omega = \sum_k P(k)p_k I_k$. Following what we did before we can take the average of both terms:

$$\partial_t \bar{I} = (\beta - \mu)\bar{I}\,, \tag{6.53}$$

so that:

$$\bar{I} = \bar{I}(0)e^{(\beta-\mu)t}. \tag{6.54}$$

Even in this case we recover the threshold $R_0 = \beta/\mu > 1$. The deterministic equations consider the reaction-diffusion processes in a mean-field fashion. The concept of invasion threshold can not emerge out of this approach. What is then the validity of the equations that we just derived? They can be used to study the evolution across the degree classes above the threshold.

6.2 Global invasion threshold in metapopulation networks with origin-destination diffusion

In the previous sections, we presented the general framework of metapopulations models. We considered homogeneous and heterogeneous metapopulations networks and different mechanism of diffusion among the different patches. We showed how the diffusion mechanism and their rates are crucial in the definition of the global invasion threshold. In the analyzed cases the approximations we made allowed us to evaluate all the interesting quantities and to have a clear understanding of the processes. The approximation related to the diffusion protocols we used so far are far away from reality. In the previous model individuals are considered as particles without home or destination during their travels. In the next sections we will use a more realistic diffusion protocol considering that individuals, at least on average, do not move randomly among cities. They have an origin and a destination. We will consider the same general framework and, in order to capture the physics of the problem, the betweenness centrality that we introduced in the Chapter 1. We will show how to solve the equations evaluating the invasion global threshold in two limits.

Let us consider now a diffusion process in which each individual in a node i travels with a probability p in random node j. Each individual will come back after reached the destination. The shortest path is selected among all the possible paths. The number of individuals that will be route in a node is proportional to the number of shortest

path among any other pair of nodes that pass through it. In these type of processes the betweenness centrality is then the natural candidate to give us the role or centrality of each node. As showed in Chapter 1 this quantity is defined as:

$$b(i) = \sum_{\substack{j,l=1,n \\ i \neq j \neq l}} \frac{\mathcal{D}_{jl}(i)}{\mathcal{D}_{jl}}, \tag{6.55}$$

where $\mathcal{D}_{jl}$ is the total number of shortest paths from j to l and $\mathcal{D}_{jl}(i)$ is the number of shortest paths from j to l that goes through i.

Let us consider that an individual of a subpopulation of degree k get some infectious disease characterized by a reproductive number $R_0 > 1$. Let us define, even in this case, D_k^0 as the number of diseased subpopulation of degree k at generation 0. In the early stage, the number of diseased subpopulations is small. We can study the evolution of this number using a tree-like approximation relating D_k^n with D_k^{n-1}. The average number of infected individuals in the class of degree k during the evolution of the epidemic is αN_k. α depends on the specific disease. Each infected individual stays in the infectious state for an average time μ^{-1}. Then the number of infected people circulating in the network after $n-1$ generations is:

$$\omega^{n-1} = \frac{p\alpha}{\mu} \sum_{k'} D_{k'}^{n-1} N_{k'}. \tag{6.56}$$

The number of infected individual that will pass through a node of class k will be a fraction of (6.56) proportional to the betweenness:

$$\gamma_k^{n-1} = \frac{b_k}{b_{tot}} \omega^{n-1}, \tag{6.57}$$

where b_{tot} is the sum of all the betweenness of the nodes. For the n^{th} generation we have:

$$D_k^n = V_k \left(1 - \frac{D_k^{n-1}}{V_k}\right) \left[1 - R_0^{-\gamma^{n-1}}\right], \tag{6.58}$$

where the first term on the right is the probability that the subpopulation is not already seeded by infected individuals and the last is the probability that the new seeded subpopulation will experience an outbreak. In the early time and for $R_0 \sim 1$ we can approximate the expression considering:

$$\frac{D_k^{n-1}}{V_k} << 1, \tag{6.59}$$

and

$$1 - R_0^{-\gamma^{n-1}} \sim (R_0 - 1)\gamma^{n-1}, \tag{6.60}$$

obtaining:

$$D_k^n = (R_0 - 1)V_k\gamma^{n-1} = (R_0 - 1)\frac{p\alpha}{\mu}V_k\frac{b_k}{b_{tot}}\sum_{k'} D_{k'}^{n-1}N_{k'}. \tag{6.61}$$

Considering at the equilibrium:

$$N_k = \frac{k}{< k >}\bar{N}, \tag{6.62}$$

where $\bar{N} = \sum_k P(k)N_k$ is the average subpopulation size, we get:

$$D_k^n = (R_0 - 1)\frac{p\alpha}{\mu}\bar{N}V_k\frac{b_k}{b_{tot}}\frac{1}{< k >}\sum_{k'} D_{k'}^{n-1}k'. \tag{6.63}$$

Let us define now $\Theta^n = \sum_k D_k^n k$, then we have:

$$\Theta^n = (R_0 - 1)\frac{p\alpha}{\mu}\bar{N}\frac{\Theta^{n-1}}{< k >}\sum_k V_k k\frac{b_k}{b_{tot}}. \tag{6.64}$$

The last term needs some more work:

$$\sum_k V_k k\frac{b_k}{b_{tot}} = \frac{V\sum_k P(k)kb_k}{V\sum_{k'} P(k')b_{k'}}. \tag{6.65}$$

Considering now $b_k \sim k^\eta$:

$$\Theta^n = (R_0 - 1)\frac{p\alpha}{\mu}\bar{N}\frac{1}{< k >}\frac{< k^{1+\eta} >}{< k^\eta >}\Theta^{n-1}. \tag{6.66}$$

We finally get the global invasion threshold:

$$R^* = (R_0 - 1)\frac{p\alpha}{\mu}\bar{N}\frac{1}{< k >}\frac{< k^{1+\eta} >}{< k^\eta >}. \tag{6.67}$$

We can write the threshold condition on the mobility rate:

$$p\bar{N} \geq \frac{< k^\eta >}{< k^{1+\eta} >}\frac{< k >}{\alpha}\mu(R_0 - 1)^{-1}. \tag{6.68}$$

These last two expression are the crucial quantities, that give us the condition for a global outbreak. It is important to remind that in metapopulations networks the condition $R_0 > 1$ for each subpopulation is not enough to judge if a finite number of subpopulations will be affected by the disease. The diffusion process must be considered, and it defines

the form and value of the invasion threshold. These arguments are valid just in the case in which:

$$\mu^{-1} \gg \bar{l}v^{-1}, \tag{6.69}$$

where $\bar{l}$ is the average distance between nodes and v is the *speed* of individuals that is fix in our case at 1 node per time step. We can get an expression in the other limit too:

$$\mu^{-1} \ll \bar{l}v^{-1}, \tag{6.70}$$

in this case each individual will be infectious for a small time window good enough to infect just the nearest neighbors. In this case the spreading can be thought as a Markovian process and the expression for the global invasion threshold will be:

$$R^* = \frac{p\alpha}{\mu}(R_0 - 1)\bar{N}\frac{< k^2 > - < k >}{< k >^2}, \tag{6.71}$$

and the mobility rate reads:

$$p\bar{N} \geq \frac{\mu}{\alpha(R_0 - 1)}\frac{< k >^2}{< k^2 > - < k >}, \tag{6.72}$$

as in the (6.31).

We can evaluate the ratio of the two different thresholds. Let us call R_1^* the (6.67) and R_2^* the (6.71):

$$\frac{R_1^*}{R_2^*} = \frac{\mu_2}{\mu_1}\frac{< k^{1+\eta} >< k >}{< k^\eta > (< k^2 > - < k >)}, \tag{6.73}$$

where we have to consider the case in which $\mu_1^{-1} \gg \mu_2^{-1}$. Setting the values $\eta = 1.1, \mu_1 = 0.01, \mu_2 = 0.5$, considering a scale-free network with exponent $\alpha = -2.1$ of $N = 10^6$ nodes with a $k_{max} \sim \sqrt{N} = 10^3$:

$$\frac{R_1^*}{R_2^*} = 65.488. \tag{6.74}$$

This is an intuitive result. If a process with a small μ^{-1} is below the global invasion threshold (i.e. $R_2^* = 0.1$) the same process with a bigger μ^{-1} can be above threshold (i.e. $R_1^* = 6.5$) and it will originate a global outbreak.

6.2.1 Comparison with numerical results

To compare the analytical insights with the numerical results we choose as substrate an uncorrelated scale free network generated according to the configuration model with $\gamma = 2.7$ and $N = 10^4$ (constrained to $k_{max} < \sqrt{N}$). First of all we tested the assumption

made in equation (6.62), in which the number of individuals N_k at nodes of degree k, at the equilibrium, is proportional to k. To do so, we started the simulation with an equal population in each node (namely $N = 10^6$ and $N_i = 10^2$), wait until the traffic equilibrium has been reached, and finally we collected the values of N_i. Figure (6.3) shows the values of N_k as function of degree k, justifying, at least for higher degrees, our assumption. In order to calculate the critical mobility rate p_c we used equation (6.68). Thus, we need to know the specific value of η in the chosen network and fix a value for R_0. To obtain an estimate for η we computed the value of the betweenness $b(i)$ for each node i and coarse grain by degree classes k. Note that to evaluate $b(i)$ of each node we decided to make a run of the simulation and register the number of individuals that pass through a link over a very long period of time, as in this way the values of $b(i)$ are more precise and near to the actual dynamics (alternatively, one could assume that the algorithmic betweenness coincides with the topological betweenness and calculate the latter directly from the topology. However, this will result in more noise). Figure (6.4) shows b_k as function of the degree classes and the fit for η gives a value of $\eta \sim 1.53$. We calculated the mean degree of the network $k = 3.8$ and then the η^{th} moment of the degree distribution obtaining $< k^{\eta} >= 11.3$ and finally the $(1+\eta)^{th}$ moment $< k^{1+\eta} >= 252$, we also decided to fix $R_0 = 2$ and the value of $\bar{N} = 100$, and substituting α in (6.68):

$$p_c = \frac{1}{\bar{N}} \frac{< k^{\eta} >}{< k^{1+\eta} >} \frac{R_0^2}{2(R_0 - 1)^2} < k > \mu = \frac{1}{100} \frac{11.3}{252} \times 2 \times 3.8 \times 0.04 = 0.0001363 \quad (6.75)$$

In Figure (6.5) we show the comparison between the numerical curve and the analytical prediction. It seems that mean-filed nicely agrees with simulation results.

All the results presented in this Chapter can be considered as propaedeutic for the next one in which a more realistic data-driven large-scale model will be introduced.
Here we have studied the general framework of metapopulations epidemic models. We described the analytical details in terms of degree block variables. In this approaches every node with the same degree is assumed to be equivalent. We showed how the system is characterized by two different epidemic thresholds defined at different scales. A local one which depends on the disease parameters values only and it determinate outbreak within the local subpopulation. A global one, called global invasion threshold, that depends on the disease parameters (the local threshold) and on the diffusion rates of the individuals. This threshold is crucial in the process and defines the range of parameters

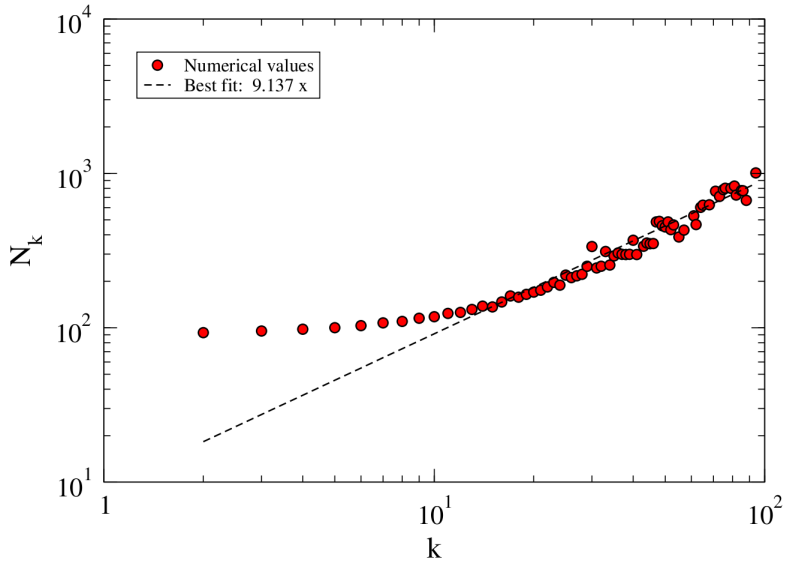

Figure 6.3: N_k as a function of the degree classes k and the best fitting curve, leading to a linear scaling.

in which a finite fraction of the subpopulations will experience an outbreak. We showed how changes in coupling between patches have critical implication for disease extinction. Homogeneous and Heterogeneous connectivity patterns have been studied. We showed how complex features has a crucial role in the dynamics of the processes, in the value and form of the global invasion threshold. In the last part of this Chapter we discussed a more realistic protocol of diffusion. We considered the fact that people do not move randomly in their trips. Origin-destination diffusion has been considered. We used betweenness centrality as a natural measure to encode the diffusion of individuals in the network considering pairs of nodes: origin and destination. We solved analytically the model in two limit finding very good agreement with the numerical simulations. Despite all the analytical and numerical success presented so far, many problems are still an open issue. For example the study of the behavior of metapopulation models with a complex internal structure in each node, or other more realistic diffusion process considering other more complicated non-Markovian processes. These are future challenge that will be considered in future works.

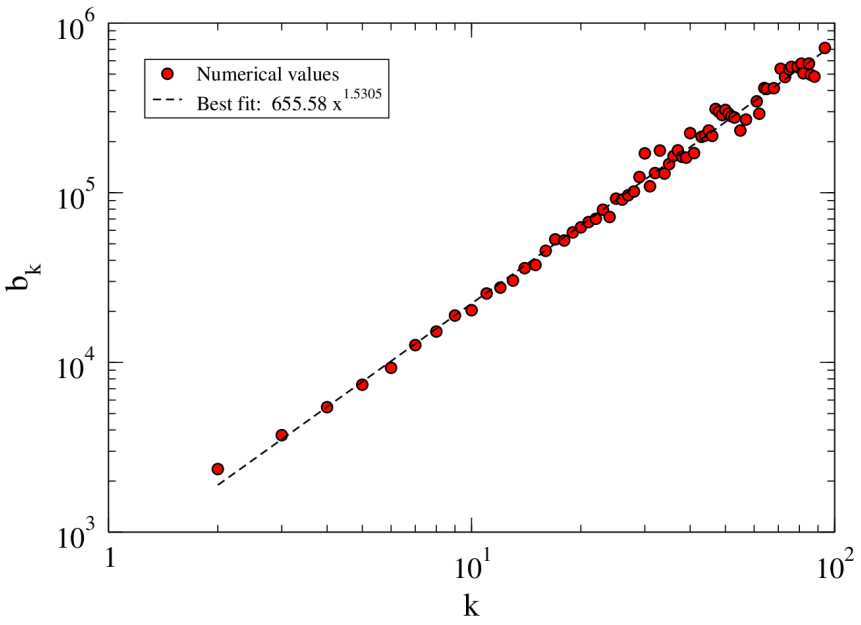

Figure 6.4: b_k as a function of the degree classes k and the best fitting curve, leading to a value of $\eta \simeq 1.53$.

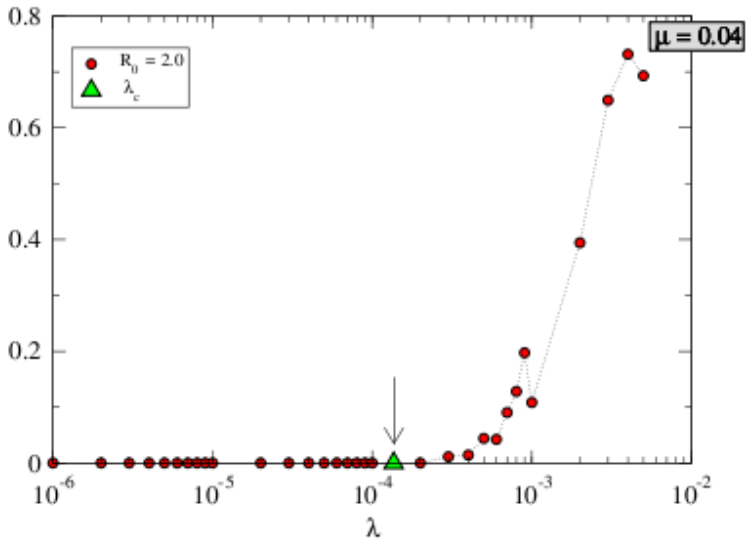

Figure 6.5: Fraction of recovered individuals in the null models as function of the mobility rate p

7

GLEaM

Physicists bring a special balance between mathematical rigor,
computational approaches and intuition for the problem.
We are artists of the approximation.
A.Vespignani

Contents

Here we present the detailed definition and data description of the *Global Epidemic and Mobility* (GLEaM) model (21). This computational model uses a structured metapopulation scheme integrating the stochastic modeling of the disease dynamics, high resolution census data and human mobility patterns at the gobal scale. We used this model during the H1N1 2009 pandemic to provide, at the global scale, epidemic forecast at different resolution of time. We will describe first the model and then we will present how we use it to model the recent pandemic.

GLEaM is a data-driven epidemic model. Many others have been proposed with an agent-based or metapopulation approach but just a few are able to consider spatio-temporal behaviour of disease at the global scale. Agent-based models are stochastic, spatially explicit, discrete time models where each agent represent a single individuals. The network of contacts is based on realistic model of the sociodemographic structure of

the population. These models are very accurate in the description of the spread of a disease, but they are based in high quality data, that world wide are not availble. Another limitation of these approach is the computational power that is needed. The combination of these two limitations have rescricted the application of agent-based models to single or few countries such the US (141; 142; 143), the UK (141), Italy (144), Thailand (145) and the entire Europe (146). Metapopulations models as we saw in Chapter 6 are based on a simple homogenous assumption inside each subpopulation. The accuracy and realism of these model relie in the ability to capture the distribution of population and the travel flows of individuals from one population to another. These approachs are then a trade off between the high realism of agent-based and computational scalability of the algorithm implementation and the relatively small amount of input data needed to feed the model that allows analysis at the worldwide scale.

7.1 The model

GLEaM integrates three diffent layers. The first one is a data layer based on high resolution population data. The second one refers to human mobility defined by the transportation and commuting networks characterizing the interactions and exchanges of individuals across subpopulations. This result in a world-wide multiscale mobility network spanning several orders of magnitude in intensity and spatio-temporal scales. The third layer is the epidemic dynamic model that defines the evolution of the infectious disease inside each subpopulations. In the next sections we will provide a detailed description of all these components.

7.1.1 Global population and subpopulations definition

The population dataset was obtained from the Web sites of the "Gridded Population of the World" and the "Global Urban-Rural Mapping" projects (147; 148), which are run by the Socioeconomic Data and Application Center (SEDAC) of Columbia University. The surface of the world is divided into a grid of cells that can have different resolution levels. Each of these cells has assigned an estimated population value.

Out of the possible resolutions, we have opted for cells of 15×15 minutes of arc to constitute the basis of our model. This corresponds to an area of each cell approximately equivalent to a rectangle of $25 \times 25\, km$ along the Equator. The dataset comprises 823 680

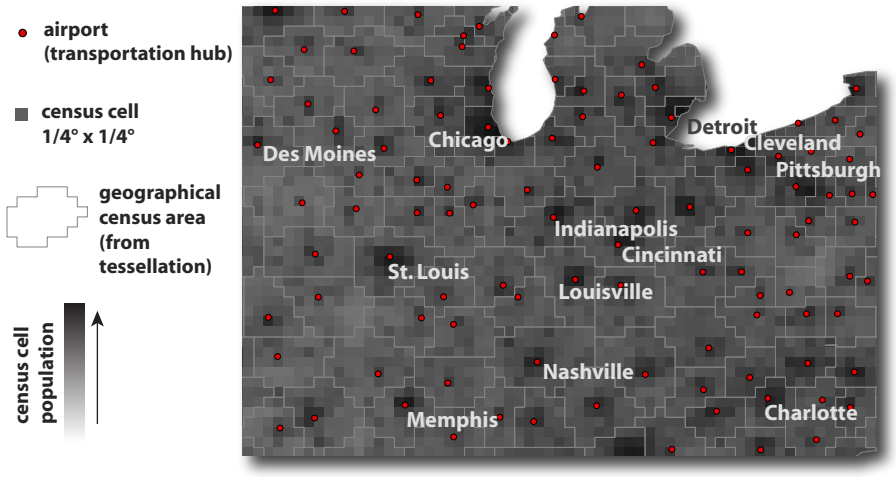

Figure 7.1: Population database and Voronoi tessellation around main tranportation hubs. The world surface is represented in a grid-like partition where each cell - corresponding to a population values - is assigned to the closest airport. Geographical census areas emerge that constitute the sub-populations of the meta-population model.

cells, of which 250 206 are populated. Since the coordinates of each cell center and those of the World Airport Network (WAN) airports are known, the distance between the cells and the airports can be calculated. We have performed a Voronoi-like tessellation (149) of the Earth surface assigning each cell to the closest airport that satisfies the following two conditions:

- Each cell is assigned to the closest airport within the same country,
- the distance between the airport and the cell cannot be longer than 200 km.

This cutoff naturally emerges from the distribution of distances between cells and closest airports, and it is introduced to avoid that in barely populated areas such as Siberia we can generate geographical census areas thousands of kilometer wide but with almost no population. It also corresponds to a reasonable upper cutoff for the ground traveling distance expected to be covered to reach an airport before traveling by plane.

Before proceeding with the tessellation, we need to take into account that some urban areas include more than one airport. For instance, London has up to six airport, Paris has two, and New York City has three. Our aim is to build a metapopulation model whose

subpopulations correspond to the geographical census areas obtained from tessellation. Inside these geographical census areas a homogeneous mixing is assumed. The groups of airports that serve the same urban area need therefore to be aggregated since the mixing within the given urban area is expected to be high and cannot be represented in terms of separated subpopulations for each of the airports serving the same city. We have searched for groups of airports located close to each other and we manually processed the identified groups of airports to select those belonging to the same urban area. The airports of the same group are then aggregated in a single "super-hub". An example with the final result of the Voronoi tessellation procedure with cells and airports can be seen in Figure (7.1). The geographical census areas become thus the basic subpopulations of our metapopulation model. Their connections will determine the geographical spreading of an hypothetical epidemic. The air transportation is already integrated in the model, but a further step must be taken in order to also include ground transportation in a realistic way.

7.1.2 World airport network

The World Airport Network (WAN) is composed of 3362 commercial airports indexed by the International Air Transport Association (IATA) that are located in 220 different countries. The database contains the number of available seats per year for each direct connection between two of these airports. The coverage of the dataset is estimated to be 99% of the global commercial traffic. The WAN can be seen as a weighted graph comprising 16 846 edges whose weight, $\omega_{j\ell}$, represents the passenger flow between airports j and ℓ. The network shows a high degree of heterogeneity both in the number of destinations per airport and in the number of passengers per connection (33; 124; 150; 151).

7.1.3 Commuting networks

Our commuting databases have been collected from the Offices of Statistics of 28 countries in the 5 populated continents. The full dataset comprehends more than 78 000 administrative regions and over five million commuting flow connections between them (21). The definition of administrative unit and the granularity level at which the commuting data are provided enormously vary from country to country. For example, most

Table 7.1: Commuting networks in each continent. Number of countries (N_c), number of administrative units (V) and inter-links between them (E) are summarized.

Continent	N_c	V	E
Europe	17	65880	4490650
North America	2	6986	182255
Latin America	4	1858	63678
Asia	3	2732	323815
Oceania	2	746	30679
Total	28	78202	5091077

European countries adhere to a practice that ranks administrative divisions in terms of geocoding for statistical purposes, the so called Nomenclature of Territorial Units for Statistics (NUTS). Most countries in the European Union are partitioned into three NUTS levels which usually range from states to provinces. The commuting data at this level of resolution is therefore strongly coarse-grained. In order to have a higher geographical resolution of the commuting datasets that could match the resolution scale of our geographical census areas, we looked for smaller local administrative units (LAU) in Europe. The US or Canada report commuting at the level of counties. However, even within a single country the actual extension, shape, and population of the administrative divisions are usually a consequence of historical reasons and can be strongly heterogeneous.

Such heterogeneity renders the efforts to define a universal law describing commuting flows likely to fail. The mobility behavior might indeed result different across countries simply due to the country specific partition of the population into administrative boundaries. In order to overcome this problem, and in particular to define a data/driven short range commuting for GLEaM, we used the geographical census areas obtained from the Voronoi tessellation as the elementary units to define the centers of gravity for the process of commuting. This allows to deal with self-similar units across the world with respect to mobility as emerged from a tessellation around main hubs of mobility and not country specific administrative boundaries. We have therefore mapped the different levels of commuting data into the geographical census areas formed by the Voronoi-like

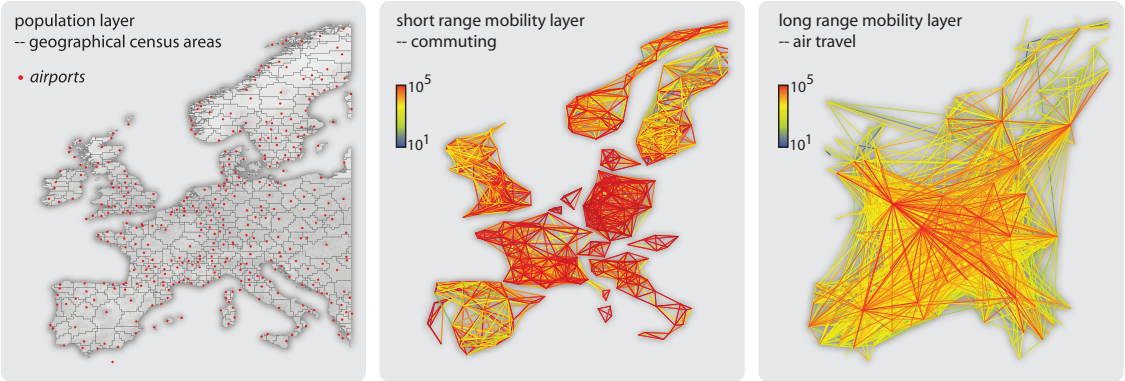

Figure 7.2: The world surface is represented in a grid-like partition where each cell – corresponding to a population value – is assigned to the closest airport. Geographical census areas emerge that constitute the subpopulations of the metapopulation model. The demographic layer is coupled with two mobility layers, the short range commuting layer and the long range air travel layer.

tessellation procedure described above. The mapped commuting flows can be seen as a second transport network connecting subpopulations that are geographically close. This second network can be overlayed to the WAN in a multi-scale fashion to simulate realistic scenarios for disease spreading. The network exhibits important variability in the number of commuters on each connection as well as in the total number of commuters per geographical census area. Being the census areas relatively homogeneous and self-similar allows us to estimate a gravity law that successfully reproduce the commuting data obtained across different continents, and provide us with estimations for the possible commuting levels in the countries for which such data is not available as in Ref. (21). In Figure (7.2) we present a skecth of these different layers, as they look in Europe.

7.1.4 Epidemic dynamic model

Each geographical census area corresponds to a subpopulation in the metapopulation model, inside which we consider a Susceptible-Latent-Infectious-Recovered (SLIR) compartmental scheme, typical of influenza-like illnesses (ILIs), where each individual has a discrete disease state assigned at each moment in time. In Figure (7.3), a diagram of the compartmental structure with transitions between compartments is shown. The contagion process, i.e. generation of new infections, is the only transition mechanism

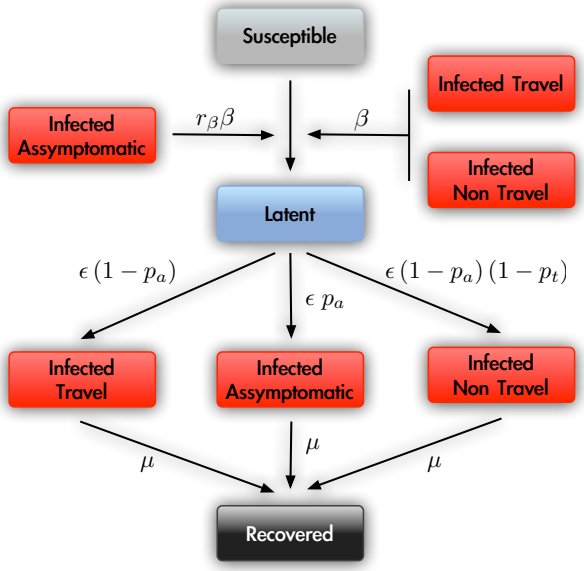

Figure 7.3: Compartmental structure of our epidemic model within each subpopulation.

which is altered by short-range mobility, whereas all the other transitions between compartments are spontaneous and remain unaffected by the commuting. The rate at which a susceptible individual in subpopulation j acquires the infection, the so called force of infection λ_j, is determined by interactions with infectious persons either in the home subpopulation j or in its neighboring subpopulations on the commuting network.

Given the force of infection λ_j in subpopulation j, each person in the susceptible compartment (S_j) contracts the infection with probability $\lambda_j \Delta t$ and enters the latent compartment (L_j), where Δt is the time interval considered. Latent individuals exit the compartment with probability $\varepsilon \Delta t$, and transit to asymptomatic infectious compartment (I_j^a) with probability p_a or, with the complementary probability $1 - p_a$, become symptomatic infectious. Infectious persons with symptoms are further divided between those who can travel (I_j^t), probability p_t, and those who are travel-restricted (I_j^{nt}) with probability $1 - p_t$. All the infectious persons permanently recover with probability $\mu \Delta t$, entering the recovered compartment (R_j) in the next time step. All transitions and corresponding rates are summarized in Table (7.2) and in Figure (7.3). In each subpopulation the variation of the number of individuals in each compartment $[m]$ can be

Table 7.2: Transitions between compartments and their rates.

Transition	Type	Rate
$S_j \to L_j$	Contagion	λ_j
$L_j \to I_j^a$	Spontaneous	εp_a
$L_j \to I_j^t$	"	$\varepsilon(1-p_a)p_t$
$L_j \to I_j^{nt}$	"	$\varepsilon(1-p_a)(1-p_t)$
$I_j^a \to R_j$	"	μ
$I_j^t \to R_j$	"	μ
$I_j^{nt} \to R_j$	"	μ

written at any given time step as:

$$X_j^{[m]}(t+\Delta t) - X_j^{[m]}(t) = \Delta X_j^{[m]} + \Omega_j([m]), \tag{7.1}$$

where the term $\Delta X_j^{[m]}$ represents the change due to the compartment transitions induced by the disease dynamics and the transport operator $\Omega_j([m])$ represents the variations due to the traveling and mobility of individuals. The latter operator takes into account the long-range airline mobility and define the minimal time scale of integration to 1 day. The mobility due to the commuting flows is taken into account by defining effective force of infections by using a time scale separation approximations as detailed in the following sections.

7.1.5 Stochastic and discrete integration of the disease dynamics

In each subpopulation j, we define an operator acting on a compartment $[m]$ to account for all the transitions out of the compartment in the time interval Δt. Each element $\mathcal{D}_j([m],[n])$ of this operator is a random variable extracted from a multinomial distribution and determines the number of transitions from compartment $[m]$ to $[n]$ occurring in Δt. The change $\Delta X_j^{[m]}$ of a compartment $[m]$ in this time interval is given by a sum over all random variables $\{\mathcal{D}_j([m],[n])\}$ as follows:

$$\Delta X_j^{[m]} = \sum_{[n]} \left\{ -\mathcal{D}_j([m],[n]) + \mathcal{D}_j([n],[m]) \right\}. \tag{7.2}$$

As a concrete example let us consider the evolution of the latent compartment. There are three possible transitions from the compartment: transitions to the asymptomatic infectious, the symptomatic traveling and the non-traveling infectious compartments. The elements of the operator acting on L_j are extracted from the multinomial distribution:

$$Pr^{Multin}(L_j(t), p_{L_j \to I_j^a}, p_{L_j \to I_j^t}, p_{L \to I_j^{nt}}), \tag{7.3}$$

determined by the transition probabilities

$$\begin{aligned} p_{L_j \to I_j^a} &= \varepsilon p_a \Delta t, \\ p_{L_j \to I_j^t} &= \varepsilon (1 - p_a) p_t \Delta t, \\ p_{L \to I_j^{nt}} &= \varepsilon (1 - p_a)(1 - p_t) \Delta t, \end{aligned} \tag{7.4}$$

and by the number of individuals in the compartment $L_j(t)$ (its size). All these transitions cause a reduction in the size of the compartment. The increase in the compartment population is due to the transitions from susceptibles into latents. This is also a random number extracted from a binomial distribution:

$$Pr^{Bin}(S_j(t), p_{S_j \to L_j}), \tag{7.5}$$

given by the chance of contagion

$$p_{S_j \to L_j} = \lambda_j \Delta t, \tag{7.6}$$

with a number of attempts given by the number of susceptibles $S_j(t)$. After extracting these numbers from the appropriate multinomial distributions, we can calculate the change $\Delta L_j(t)$ as:

$$\Delta L_j(t) = L_j(t+1) - L_j(t) = -\left[\mathcal{D}_j(L, I^a) + \mathcal{D}_j(L, I^t) + \mathcal{D}_j(L, I^{nt})\right] + \mathcal{D}_j(S, L). \tag{7.7}$$

7.1.6 The integration of the transport operator

The transport operator is defined by the airline transportation data and sets the integration time scale to 1 day. The number of individuals in the compartment $[m]$ traveling from the subpopulation j to the subpopulation ℓ is an integer random variable, in that each of the X_j potential travelers has a probability $p_{j\ell} = w_{j\ell}/N_j$ to go from j to ℓ. In

each subpopulation j the numbers of individuals $\xi_{j\ell}$ traveling on each connection $j \to \ell$ at time t define a set of stochastic variables which follows the multinomial distribution:

$$P(\{\xi_{j\ell}\}) = \frac{X_j^{[m]}!}{(X_j^{[m]} - \sum_\ell \xi_{j\ell})! \prod_\ell \xi_{j\ell}!}(1 - \sum_\ell p_{j\ell})^{(X_j^{[m]} - \sum_\ell \xi_{j\ell})} \prod_\ell p_{j\ell}^{\xi_{j\ell}}, \tag{7.8}$$

where $(1 - \sum_\ell p_{j\ell})$ is the probability of not traveling, and $(X_j^{[m]} - \sum_\ell \xi_{j\ell})$ identifies the number of non traveling individuals of the compartment $[m]$. We use standard numerical subroutines to generate random numbers of travelers following these distributions. The transport operator in each subpopulation j is therefore written as:

$$\Omega_j([m]) = \sum_\ell (\xi_{\ell j}(X_\ell^{[m]}) - \xi_{j\ell}(X_j^{[m]})), \tag{7.9}$$

where the mean and variance of the stochastic variables are $\langle \xi_{j\ell}(X_j^{[m]}) \rangle = p_{j\ell} X_j^{[m]}$ and $\mathrm{Var}(\xi_{j\ell}(X_j^{[m]})) = p_{j\ell}(1 - p_{j\ell}) X_j^{[m]}$. Direct flights as well as connecting flights up to two-legs flights can be considered. It is worth remarking that on average the airline network flows are balanced so that the subpopulation N_j are constant in time, e.g. $\sum_{[m]} \Omega_j([m]) = 0$.

7.1.7 Time-scale separation and the integration of the commuting flows

The Global Epidemic and Mobility (GLEaM) modeler combines the infection dynamics with long- and short-range human mobility. Each of these dynamical processes operates at a different time scale. For ILI there are two important intrinsic time scales, given by the latency period ε^{-1} and the duration of infectiousness μ^{-1}, both larger than 1 *day*. The long-range mobility given by the airline network has a time scale of the order of 1 *day*, while the commuting takes place in a time scale of approx. $\tau^{-1} \sim 1/3$ *day*. The explicit implementation of the commuting in the model thus requires a time interval shorter than the minimal time of airline transportation. To overcome this problem, we use a time-scale separation technique, in which the short-time dynamics is integrated into an effective force of infection in each subpopulation.

We start by considering the temporal evolution of subpopulations linked only by commuting flows and evaluate the relaxation time to an equilibrium configuration. Consider the subpopulation j coupled by commuting to other n subpopulations. The commuting rate between the subpopulation j and each of its neighbors i will be given by σ_{ji}. The

return rate of commuting individuals is set to be τ. Following the work of Sattenspiel and Dietz (152), we can divide the individuals original from the subpopulation j, N_j, between $N_{jj}(t)$ who are from j are located in j at time t and those, $N_{ji}(t)$, that are from j are located in a neighboring subpopulation i at time t. Note that by consistency:

$$N_j = N_{jj}(t) + \sum_i N_{ji}(t). \tag{7.10}$$

The rate equations for the subpopulation size evolution are then:

$$\begin{aligned} \partial_t N_{jj} = & \quad -\textstyle\sum_i \sigma_{ji} N_{jj}(t) + \tau \sum_i N_{ji}(t), \\ \partial_t N_{ji} = & \quad \sigma_{ji} N_{jj}(t) - \tau N_{ji}(t). \end{aligned} \tag{7.11}$$

By using condition (7.10), we can derive the closed expression

$$\partial_t N_{jj} + (\tau + \sigma_j) N_{jj}(t) = N_j \tau, \tag{7.12}$$

where σ_j denotes the total commuting rate of population j, $\sigma_j = \sum_i \sigma_{ji}$. $N_{jj}(t)$ can be expressed as:

$$N_{jj}(t) = e^{-(\tau+\sigma_j)t} \left(C_{jj} + N_j \tau \int_0^t e^{(\tau+\sigma_j)s} ds \right), \tag{7.13}$$

where the constant C_{jj} is determined from the initial conditions, $N_{jj}(0)$. The solution for $N_{jj}(t)$ is then:

$$N_{jj}(t) = \frac{N_j}{(1+\sigma_j/\tau)} + \left(N_{jj}(0) - \frac{N_j}{(1+\sigma_j/\tau)} \right) e^{-\tau(1+\sigma_j/\tau)t}. \tag{7.14}$$

We can similarly solve the differential equation for the time evolution of $N_{ji}(t)$:

$$\begin{aligned} N_{ji}(t) &= \frac{N_j \sigma_{ji}/\tau}{(1+\sigma_j/\tau)} - \frac{\sigma_{ij}}{\sigma_j} \left(N_{jj}(0) - \frac{N_j}{(1+\sigma_j/\tau)} \right) e^{-\tau(1+\sigma_j/\tau)t} \\ &+ \left[N_{ji}(0) - \frac{N_j \sigma_{ji}/\tau}{(1+\sigma_j/\tau)} + \frac{\sigma_{ij}}{\sigma_j} \left(N_{jj}(0) - \frac{N_j}{(1+\sigma_j/\tau)} \right) \right] e^{-\tau t}. \end{aligned} \tag{7.15}$$

The relaxation to equilibrium of N_{jj} and N_{ji} is thus controlled by the characteristic time $[\tau\,(1+\sigma_j/\tau)]^{-1}$ in the exponentials. Such term is dominated by $1/\tau$ if the relation $\tau \gg \sigma_j$ holds. In our case, $\sigma_j = \sum_i \omega_{ji}/N_j$, that equals the daily total rate of commuting for the population j. Such rate is always smaller than one since only a fraction of the local population is commuting, and it is typically much smaller than $\tau \simeq 3 - 10\, day^{-1}$. Therefore the relaxation characteristic time can be safely approximated by $1/\tau$. This

time is considerably smaller than the typical time for the air connections of one day and hence our approximation of considering the subpopulations $N_{jj}(t)$ and $N_{ji}(t)$ as relaxed to their equilibrium values:

$$N_{jj} = \frac{N_j}{1+\sigma_j/\tau} \quad \text{and} \quad N_{ji} = \frac{N_j \sigma_{ji}/\tau}{1+\sigma_j/\tau}, \tag{7.16}$$

is reasonable. This approximation, originally introduced by Keeling and Rohani (153), allows us to consider each subpopulation j as having an effective number of individuals N_{ji} in contact with the individuals of the neighboring subpopulation i. In practice, this is similar to separate the commuting time scale from the other time scales in the problem (disease dynamics, traveling dynamics, etc.). While the approximation holds exactly only in the limit $\tau \to \infty$, it is good enough as long as τ^{-1} is much smaller than the typical transition rates of the disease dynamics. In the case of ILIs, the typical time scale separation between τ and the compartments transition rates is close to one order of magnitude or even larger. The equations (7.17) can be then generalized in the time scale separation regime to all compartments $[m]$ obtaining the general expression:

$$X_{jj}^{[m]} = \frac{X_j^{[m]}}{(1+\sigma_j/\tau)} \text{ and } X_{ji}^{[m]} = \frac{X_j^{[m]}}{(1+\sigma_j/\tau)} \sigma_{ji}/\tau, \tag{7.17}$$

where $\sigma_j = \sum_{i \in v(j)} \sigma_{ji}$ denotes the total commuting rate of j. Whereas $X_{jj}^{[m]} = X_j^{[m]}$ and $X_{ji}^{[m]} = 0$ for all the other compartments which are restricted from traveling. These expressions will be used to obtain the effective force of infection taking into account the interactions generated by the commuting flows.

7.1.8 Effective force of infection

The force of infection λ_j that a susceptible population of a subpopulation j sees can be decomposed into two terms: λ_{jj} and λ_{ji}. The component λ_{jj} refers to the part of the force of infection whose origin is local in j. While λ_{ji} indicates the force of infection acting on susceptibles of j during their commuting travels to a neighboring subpopulation i. The effective force of infection can be estimated by summing these two terms weighted by the probabilities of finding a susceptible from j in the different locations, S_{jj}/S_j and S_{ji}/S_j, respectively. Using the time-scale separation approximation that establishes the equilibrium populations in equation (7.17), we can write:

$$\lambda_j = \frac{\lambda_{jj}}{1+\sigma_j/\tau} + \sum_i \frac{\lambda_{ji}\sigma_{ji}/\tau}{1+\sigma_j/\tau}. \tag{7.18}$$

We will focus now on the calculation of each term of the previous expression. The force of infection occurring in a subpopulation j is due to the local infectious persons staying at j or to infectious individuals from a neighboring subpopulation i visiting j and so we can write:

$$\lambda_{jj} = \frac{\beta_j}{N_j^*}\left[I_{jj}^{nt} + I_{jj}^{t} + r_\beta I_{jj}^{a} + \sum_i \left(I_{ij}^{nt} + I_{ij}^{t} + r_\beta I_{ij}^{a}\right)\right], \tag{7.19}$$

where β_j is introduced to account for the seasonality in the infection transmission rate (if the seasonality is not considered, it is a constant), and N_j^* stands for the total effective population in the subpopulation j. By definition, $I_{jj}^{nt} = I_j^{nt}$ and $I_{ji}^{nt} = 0$ for $j \neq i$. If we use the equilibrium values of the other infectious compartments (see equation (7.17)) we obtain:

$$\lambda_{jj} = \frac{\beta_j}{N_j^*}\left[I_j^{nt} + \frac{I_j^t + r_\beta I_j^a}{1+\sigma_j/\tau} + \sum_i \frac{I_i^t + r_\beta I_i^a}{1+\sigma_i/\tau}\sigma_{ij}/\tau\right]. \tag{7.20}$$

The derivation of λ_{ji} follows from a similar argument yielding:

$$\lambda_{ji} = \frac{\beta_i}{N_i^*}\left[I_{ii}^{nt} + I_{ii}^{t} + r_\beta I_{ii}^{a} + \sum_{\ell\in\upsilon(i)} \left(I_{\ell i}^{nt} + I_{\ell i}^{t} + r_\beta I_{\ell i}^{a}\right)\right], \tag{7.21}$$

where $\upsilon(i)$ represents the set of neighbors of i, and therefore the terms under the sum are due to the visits of infectious individuals from the subpopulations ℓ, neighbors of i, to i. By plugging the equilibrium values of the compartment into the above expression, we obtain:

$$\lambda_{ji} = \frac{\beta_i}{N_i^*}\left[I_i^{nt} + \frac{I_i^t + r_\beta I_i^a}{1+\sigma_i/\tau} + \sum_{\ell\in\upsilon(i)} \frac{I_\ell^t + r_\beta I_\ell^a}{1+\sigma_\ell/\tau}\sigma_{\ell i}/\tau\right]. \tag{7.22}$$

Finally, in order to have an explicit form of the force of infection we need to evaluate the effective population size N_j^* in each subpopulation j, i.e., the actual number of people actually staying at the location j. The effective population is $N_j^* = N_{jj} + \sum_i N_{ij}$, that in the time-scale separation approximation reads:

$$N_j^* = I_j^{nt} + \frac{N_j - I_j^{nt}}{1+\sigma_j/\tau} + \sum_i \frac{N_i - I_i^{nt}}{1+\sigma_i/\tau}\sigma_{ij}/\tau. \tag{7.23}$$

Note that in these equations all the terms with compartments have an implicit time dependence. By inserting λ_{jj} and λ_{ji} into equation (7.18), it can be seen that the expression for the force of infection includes terms of zeroth, first and second order on the

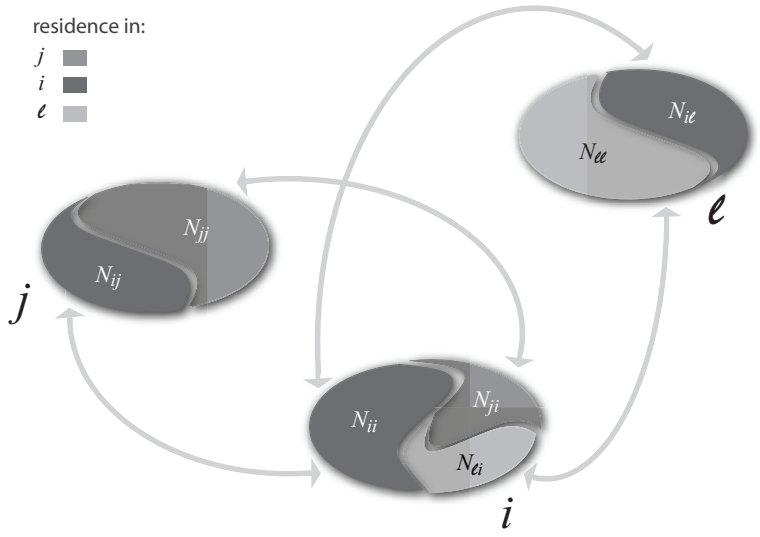

Figure 7.4: Schematic representation of the subdivision of the population in each geographical census area. The population in each geographical census area is divided into partial populations N_{xy}, where x represents the subpopulation of residence and y represents the subpopulation of the actual location at time t. Three subpopulations are shown – i, j, ℓ – to represent the various contributions to the force of infection (see equation (7.18))

commuting ratios (i.e., σ_{ij}/τ). These three term types have a straightforward interpretation: The zeroth order terms represent the usual force of infection of the compartmental model with a single subpopulation. The first order terms account for the effective contribution generated by neighboring subpopulations with two different sources: Either susceptible individuals of subpopulation j having contacts with infectious individuals of neighboring subpopulations i, or infectious individuals of subpopulations i visiting subpopulation j. The second order terms correspond to an effective force of infection generated by the contacts of susceptible individuals of subpopulation j meeting infectious individuals of subpopulation ℓ (neighbors of i) when both are visiting subpopulation i (see Figure (7.4)). This last term is very small in comparison with the zeroth and first order terms, typically around two order of magnitudes smaller, and in general can be neglected.

Algorithm 1 Generic GLEaM program flow.

Parse model file
Load data input files:
    population database
    commuting
    flight networks

foreach timestep t:
do
    Flight connections (See Alg. 2)
    Infect (See Alg. 3)
    Aggregate results for each detail level.
done

Generate final output

7.1.9 Seasonality modeling

To model seasonal variations we follow the approach of Cooper *et al* (154) and scale the basic reproduction ratio R_0 by a seasonal function, $s_i(t)$,

$$s_i(t) = \left[\left(1 - \frac{R_{min}}{R_{max}}\right) \sin\left(\frac{2\pi}{365}(t - t_{max,i}) + \frac{\pi}{2}\right) + 1 + \frac{R_{min}}{R_{max}}\right]\frac{1}{2} \quad , \tag{7.24}$$

where i stands for the North or South hemispheres. This function is identically equal to 1.0 in the tropical regions. $t_{max,i}$ is the time corresponding to the maximum seasonal effect, Jan 15 in the North and six months later in the South. Seasonality has a dual effect, it increases the value of R_0 up to $R_{max} = \alpha_{max} R_0$ with $\alpha_{max} \equiv 1.1$ (134) and reduces it down to $R_{min} = \alpha_{min} R_0$.

7.1.10 Algorithms, the simulator and its implementation

The GLEaM is a intricate model. Its implementation is not trivial. Here we will discuss step by step the implementation. The coding is done in a modular way, each module is build to performs a single function. In Algorithm (1) we report the general program flow of a GLEaM typical run.

Algorithm 2 Long distance mobility.

foreach city i:
do

  foreach neighbor $j \in v(i)$:
  do
    Calculate traffic:
      $\tilde{\omega}_{ij} = \omega_{ij} [\alpha + \eta (1 - \alpha)]$
    Traveling probability:
      $p_{ij} = \frac{\tilde{\omega}_{ij}}{N_i}$
  done

  distribute travelers among neighbors
  updated population matrix
end

Flights: long distance travel

As we said in the previous sections, our basic time scale is a day. At the start of the time step we use the flight network for the diffusion of people. The travel is considered as instantaneous, no transitions are possible on route. Travelers arrives at destination at the beginning of the day, so they have a full day chance of interact with others individuals. The probability of traveling are not fixed. We apply noise to α the occupancy rate of flights with a stochastic random variable η uniformly distributed in the interval $[-1, 1]$. The details are represented in the Algorithm (2).

Compartment transitions

The GLEaM is able to consider any compartmental model for the epidemic dynamics. The model definition is in fact one part of the input files and it is processed to generate a directed multigraph, where each node is a compartment and each edge a transition. The edges contain as attribute all the information necessary to calculate the transition probabilities, and are directly used as argument of a multinomial function in order to calculate the number of individuals leaving one compartment for another. The type of transitions are the same analyzed in the Chapter 2: interaction $A + B \rightarrow 2B$ or spontaneous $A \rightarrow B$. As we described in the previous section the commuting between

Algorithm 3 Compartment transitions.

foreach city i:
do
calculate effective populations due to commuting

foreach initial compartment x:
do
Update transition probability to compart. y using equation (7.20) and equation (7.22).
For seasonal transitions, scale transition rate by $s(t)$ (equation (7.24))
done

Move population between compartments using a multinomial
done

closest basins is encoded in the effective size of populations. As show in Algorithm (3) for each basin the first step is to evaluate the effect of commuting in terms of effective population. After this step the transitions between compartments are evaluated.

Post-processing

The GLEaM is a stochastic model. All the processes described are steps of a single run, we usually use order of 10^3 runs. In each of those the number of information that are written in a file are arbitrary decided dependently of the necessity. The amount of data could be then huge, size of compartments for each time step, number of transitions etc.. These informations could be at different resolution: census area, country, region, continent, hemisphere, globe. For this reason, the final step after each simulated day is a partial aggregation of the results in order to accelerate the post-process analysis and to reduce the amount of data wrote in each run. The post-processes take these partial aggregate output and generate the analysis, figures and eventually animation that are needed.
The full simulation process is schematically shown in Figure (7.5)

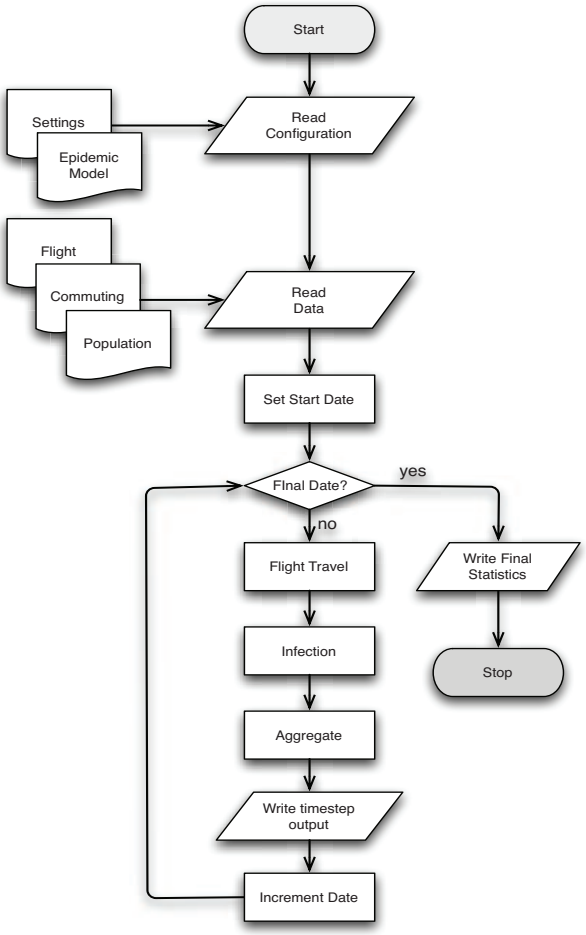

Figure 7.5: Full illustration of the procedure used for the GLEaM simulation engine. The left column represents input databases and the right column the data structures that are generated. Program flow occurs along the center. The three steps in the center box are repeated for each simulated day.

7.2 H1N1 pandemic

Here we present an estimate of the reproduction number R_0 of the H1N1 epidemic based on knowledge of human mobility patterns. We use the GLEaM (21; 155) for the worldwide evolution of the pandemic and perform a maximum likelihood analysis of the parameters against the actual chronology of newly infected countries. The method is computationally intensive as it involves a Monte Carlo generation of the distribution of arrival time of the infection in each country based on the analysis of 10^6 worldwide simulations of the pandemic evolution with the GLEaM model. We found the best estimate $R_0 = 1.75$ (95% confidence interval (CI) 1.64 to 1.88) for the basic reproductive number. Correlation analysis allows the selection of the most probable seasonal behavior based on the observed pattern, leading to the identification of plausible scenarios for the unfolding of the pandemic and the estimate of pandemic activity peaks in the different hemispheres. We provide estimates for the number of hospitalizations and the attack rate for the next wave as well as an extensive sensitivity analysis on the disease parameter values. We also studied the effect of systematic therapeutic use of antiviral drugs on the epidemic timeline and an estimation of the initial number of cases in Mexico. The analysis showed the potential for an early epidemic peak occurring in October 2009 in the Northern hemisphere, as confirmed later by the data from the surveillance, unfortunately before the large-scale vaccination campaigns was carried out. The baseline results refer to a worst-case scenario in which additional mitigation policies are not considered.

7.2.1 Background and the Epidemic Timeline

Beginning April, 2009, the world experienced its latest global pandemic outbreak originated in Mexico. It spreads quickly to many countries in months and on June 11th, 2009, the World Health Organization has officially raised the phase of pandemic alert to level 6. As of July 19th, 2009, 137,232 cases of the new H1N1 influenza strain have been officially confirmed in 142 different countries, and during the summer of 2009, the pandemic unfolding in the Southern hemisphere was under scrutiny to gain insights about the next winter wave in the North. A major challenge was given by the need to estimate the virus transmission potential and to assess its dependence on seasonality aspects in order to use numerical models capable to project the spatio-temporal pattern of the

pandemic.

Estimating the transmission potential of a newly emerging virus is crucial when planning for adequate public health interventions to mitigate its spread and impact, and to forecast the expected epidemic scenarios through sophisticate computational approaches (141; 142; 151; 156). With the recent outbreak of the new influenza A(H1N1) strain having reached pandemic proportions, the investigation of the influenza situation worldwide might provide the key to the understanding of the transmissibility observed in different regions and to the characterization of possible seasonal behavior. During the early phase of an outbreak, this task is hampered by inaccuracies and incompleteness of available information. Reporting is constrained by the difficulties in confirming large numbers of cases through specific tests and serological analysis. The cocirculation of multiple strains, the presence of asymptomatic cases that go undetected, the impossibility to monitor mild cases that do not seek health care and the possible delays in diagnosis and reporting, all worsen the situation. Early modeling approaches and statistical analysis show that the number of confirmed cases by the Mexican authorities during the early phase was underestimated by a factor ranging from one order of magnitude (157) to almost three (113). The Centers for Disease Control (CDC) in the US estimate a 5% to 10% case detection, similar to other countries facing large outbreaks, with expected heterogeneities due to different surveillance systems. Even within the same country, the setup of enhanced monitoring led to improved notification with respect to the earlier phase of the pandemic, later relaxed as reporting requirements changed (158).

By contrast, the effort put in place by the World Health Organization (WHO) and health protection agencies worldwide provided an unprecedented amount of data and, at last, the possibility of following in real time the pandemic chronology on the global scale. In particular, the border controls and the enhanced surveillance aimed at detecting the first cases reaching uninfected countries appear to provide more reliable and timely information with respect to the raw count of cases as local transmission occurs, and this data has already been used for early assessment of the number of cases in Mexico (157). Moreover, data on international passenger flows from Mexico was found to display a strong correlation with confirmed H1N1 importations from Mexico (159).

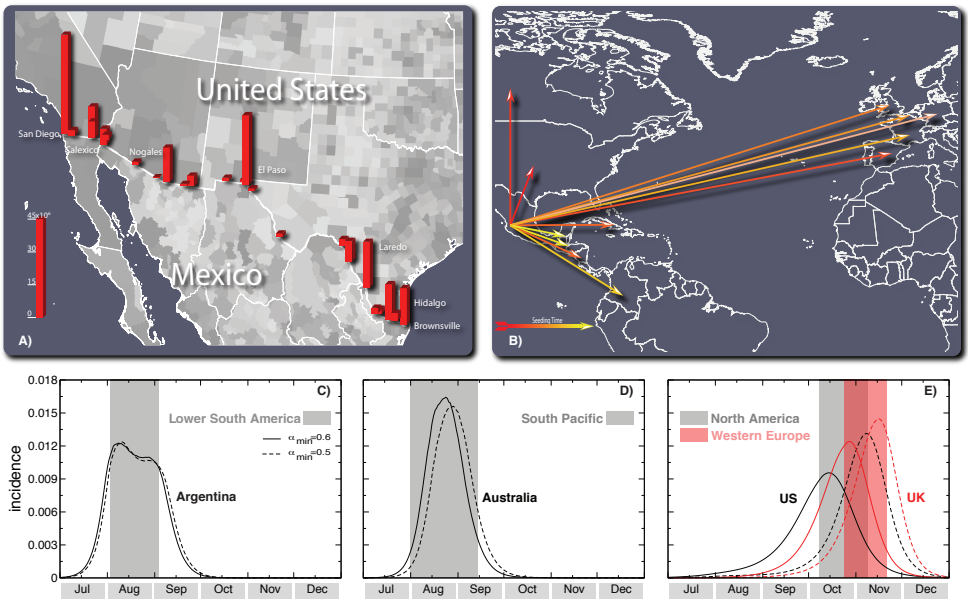

Figure 7.6: Illustration of the model's initialization and the results for the activity peaks in three geographical areas. (a) Intensity of the commuting between US and Mexico at the border of the two countries. (b) The 12 countries infected from Mexico used in the Monte Carlo likelihood analysis. The color scale of the arrows from red to yellow indicates the time ordering of the epidemic invasion. Panels (c), (d) and (e) show the daily incidence in Lower South America, South Pacific and North America/Western Europe, respectively. The shaded area indicates the 95% confidence interval (CI) of the peak time in the corresponding geographical region. The median incidence profiles of selected countries are shown for the two values defining the best-fit seasonality scaling factor interval.

7.2.2 Long-term Predictions: model and parameters

We used the classic influenza-like illness compartmentalization in which each individual is classified by a discrete state such as susceptible, latent, infectious symptomatic, infectious asymptomatic or permanently recovered/removed (76; 160). The model therefore assumes that the latent period is equivalent to the incubation period and that no secondary transmissions occur during the incubation period (see Figure (7.3) for a detailed description of the compartmentalization). As we said all transitions are modeled through binomial and multinomial processes to preserve the discrete and stochastic nature of the processes. Asymptomatic individuals are considered as a fraction $p_a = 33\%$ of the infectious individuals (161) generated in the model and assumed to infect with a relative infectiousness of $r_\beta = 50\%$ (157; 160; 162). Change in traveling behavior after the onset of symptoms is modeled with the probability $1 - p_t$, set to 50%, that individuals would stop traveling when ill (160). The spreading rate of the disease is ultimately governed by the basic reproduction number R_0. Once the disease parameters and initial conditions based on available data are defined, GLEaM allows the generation of stochastic realizations of the worldwide unfolding of the epidemic, with mobility processes entirely based on real data. The model generates in silico epidemics for which we can gather information such as prevalence, morbidity, number of secondary cases, number of imported cases and other quantities for each subpopulation and with a time resolution of 1 day. While global models are generally used to produce scenarios in which the basic disease parameters are defined from the outset, here we use the model to provide a maximum likelihood estimate of the transmission potential by finding the set of disease parameters that best fit the data on the arrival time of cases in different countries worldwide. The projections for the winter season in the northern hemisphere are also assuming that there will be no mutation of the virus with respect to the spring/summer of 2009. Furthermore, while at the moment of our analysis the novel H1N1 influenza was accounting for 75% of the influenza cases worldwide, the model did not consider the cocirculation of different influenza strains and cannot provide information on cocirculation data.

The initial conditions of the epidemic are defined by setting the onset of the outbreak near La Gloria in Mexico on 18 February 2009, as reported by official sources (163) and analogously to other works (157). We tested different localizations of the first cases in census areas close to La Gloria without observing relevant variations with respect to the

observed results. We also performed sensitivity analysis on the starting date by selecting a seeding date anticipated or delayed by 1 week with respect to the date available in official reports (163). The arrival time of infected individuals in the countries seeded by Mexico is clearly a combination of the number of cases present in the originating country (Mexico) and the mobility network, both within Mexico and connecting Mexico with countries abroad. For this reason we integrated into our model the data on Mexico-US border commuting (see Figure (7.6)-a), which could be relevant in defining the importation of cases in the US, along with Mexican internal commuting patterns that are responsible for the diffusion of the disease from rural areas as La Gloria to transportation hubs such as Mexico City. In addition, we used a time-dependent modification of the reproductive number in Mexico as in (113) to model the control measures implemented in the country starting 24 April and ending 10 May, as those might affect the spread to other countries. In order to ascertain the effect of seasonality on the observed pattern, we explored different seasonality schemes. The seasonality is modeled by a standard forcing that rescales the value of the basic reproductive number into a seasonally rescaled reproductive number, $R(t)$, depending on time. The seasonal rescaling is time and location dependent by means of a scaling multiplicative factor generated by a sinusoidal function with a total period of 12 months oscillating in the range α_{min} to α_{max}, with $\alpha_{max} = 1.1$ days (sensitivity analysis in the range 1.0 to 1.1) and α_{min} a free parameter to be estimated (154). The rescaling function is in opposition in the Northern and Southern hemispheres. No rescaling is assumed in the Tropics. The value of R_0 reported in the tables and the definition of the baseline is the reference value in the Tropics. In each subpopulation the $R(t)$ relative to the corresponding geographical location and time of the year is used in the simulations.

We have defined a Monte Carlo likelihood analysis for the assessment of the seasonal transmission potential of the new A(H1N1) influenza based on the analysis of the chronology of case detection in affected countries at the early stage of the epidemic. This method allows the use of data coming from the border controls and the enhanced surveillance aimed at detecting the first cases reaching uninfected countries. This data is, in principle, more reliable than the raw count of cases provided by countries during the evolution of the epidemic. The procedure provided the necessary input to the large-scale computational model for the analysis of the unfolding of the pandemic. The seasonal transmission potential of the H1N1 strain is assessed in a two-step process that first estimates the

reproductive number in the Tropics region, where seasonality is assumed not to occur, by focusing on the early international seeding by Mexico, and then estimates the degree of seasonal dumping factor by examining a longer time period of international spread to allow for seasonal changes. The estimation of the reproductive number is performed through a maximum likelihood analysis of the model fitting the data of the early chronology of the H1N1 epidemic. Given a set of values of the disease parameters, we produced $2 \cdot 10^3$ stochastic realizations of the pandemic evolution worldwide for each R_0 value. Our model explicitly takes into account the class of symptomatic and asymptomatic individuals and allows the tracking of the importation of each symptomatic individual and of the onset of symptoms of exposed individuals transitioning to the symptomatic class, as observables of the simulations. This allows us to obtain numerically with a Monte Carlo procedure the probability distribution $P_i(t_i)$ of the importation of the first infected individual or the first occurrence of the onset of symptoms for an individual in each country i at time t_i. Asymptomatic individuals do not contribute to the definition of t_i. With the aim of working with conditional independent variables we restrict the likelihood analysis to 12 countries seeded from Mexico (see Figure (7.6)-b) and for which it is possible to know with good confidence the onset of symptoms and/or the arrival date of the first detected case (see Table (7.3)). This allows us to define a likelihood function:

$$\mathcal{L} = \prod_i P_i(t_i^*), \tag{7.25}$$

where t_i^* is the empirical arrival time from the H1N1 chronological history in each of the selected countries. Maximizing this function, after fixing the values of the epidemiological and seasonality parameters (ϵ, μ, α_{min}), we obtain an estimation of the basic reproductive number. This methodology assumes the prompt detection of symptomatic cases at the very beginning of the outbreak in a given country, and for this reason we have also provided a sensitivity analysis accounting for a late/missed detection of symptomatic individuals as reported in the next section. The transmission potential is estimated as the value of R_0 that maximizes the likelihood function $\mathcal{L}$, for a given set of values of the disease parameters. In Table (7.5) we report the reference values assumed for some of the model parameters and the range explored with the sensitivity analysis. So far there are no precise clinical estimates of the basic model parameters ϵ and μ defining the inverse average exposed and infectious time durations (164; 165; 166). The generation interval

G_t (167; 168) used in the literature is based on the early estimate of (157) and values obtained for previous pandemic and seasonal influenza (151; 160; 161; 162; 169; 170), with most studies focusing on values ranging from 2 to 4 days (157; 171; 172; 173). We have therefore assumed a short exposed period value $\epsilon^{-1} = 1.1$ as indicated by early estimates (157) and compatible with recent studies on seasonal influenza (161; 174) and performed a sensitivity analysis for values as large as $\epsilon^{-1} = 2.5$ days. The maximum likelihood procedure is performed by systematically exploring different values of the generation time aimed at providing a best estimate and confidence interval for G_t, along with the estimation of the maximum likelihood value of R_0.
The major problem in the case of projections on an extended time horizon is the seasonality effect that in the long run is crucial in determining the peak of the epidemic. In order to quantify the degree of seasonality observed in the current epidemic, we estimate the minimum seasonality scaling factor α_{min} of the sinusoidal forcing by extending the chronology under study and analyzing the whole data set composed of the arrival dates of the first infected case in the 93 countries affected by the outbreak as of 18 June. We studied the correlation between the simulated arrival time by country and its corresponding empirical value, by measuring the regression coefficient between the two datasets. Given the extended time frame under observation, the arrival times considered in this case are expected to provide a signature of the presence of seasonality. They included the seeding of new countries from outbreaks taking place in regions where seasonal effects might occur, as for example in the US or in the UK. For the simulated arrival times we have considered the median and 95% confidence interval (CI) emerging from the $2 \cdot 10^3$ stochastic runs. The regression coefficient is found to be sensitive to variations in the seasonality scaling factor, allowing discrimination of the α_{min} value that best fits the real epidemic. The full exploration of the phase space of epidemic parameters and seasonality scenarios required data from 10^6 simulations; the equivalent of 2 million minutes of PowerPC 970 2.5 GHz CPU time.

7.2.3 Results

Table (7.5) reports the results of the maximum likelihood procedure and of the correlation analysis on the arrival times for the estimation of α_{min}. In the following we consider as the baseline case the set of parameters defined by the best estimates: $G_t = 3.6$

Country	Onset of symptoms	Flight arrival	Confirmed on
United States	March 28 (175)	–	April 21 (175)
Canada	April 11 (176)	April 8 (177)	April 23 (178)
El Salvador	–	April 19 (179)	May 3 (180)
United Kingdom	April 24 (181)	April 21 (182)	April 27 (178)
Spain	April 25 (183)	April 22 (184)	April 27 (178)
Cuba	–	April 25 (185)	May 13 (178)
Costa Rica	April 25 (186)	April 25 (186)	May 2 (178)
Netherlands	–	April 27 (187)	April 30 (187)
Germany	April 28 (188)	–	April 29 (178)
France	–	–	May 1 (189)
Guatemala	May 1 (190)	–	May 5 (191)
Colombia	–	–	May 3 (192)

Table 7.3: The day of onset of symptoms, flight arrival and day of official confirmation of the first confirmed case in 12 countries seeded by Mexico are reported.

days, $\mu^{-1} = 2.5$ days, $R_0 = 1.75$. The best estimates for G_t and R_0 are higher than those obtained in early findings but close to subsequent analysis on local outbreaks (171; 172; 173). The R_0 we report is the reference value for Mexico and the tropical region, whereas in each country we have to consider the $R(t)$ due to the seasonality rescaling depending on the time of the year, as shown in Table (7.6). This might explain the lower values found in some early analysis in the US. The transmission potential emerging from our analysis is close to estimates for previous pandemics (131; 193). In supplementary informations of the paper (16) we provide supplementary tables for the full sensitivity analysis concerning the assumptions used in the model. Results show that larger values of the generation interval provide increasing estimates for R_0. Fixing the latency period to $\epsilon^{-1} = 1.1$ days and varying the mean infectious period in the plausible range 1.1 to 4.0 days yields corresponding maximum likelihood estimates for R_0 in the range 1.4 to 2.1. Variations in the latency period from $\epsilon^{-1} = 1.1$ to $\epsilon^{-1} = 2.5$ days provide corresponding best estimates for R_0 in the range 1.9 to 2.3, if we assume an infectious period of 3 days. We tested variations of the compartmental model parameters p_a, and p_t up to 20% and explored the range $r_\beta = 20\%$ to 80%, and sensitivity on

the value of the maximum seasonality scaling factor α_{max} in the range 1.0 to 1.1. The obtained estimates lie within the confidence intervals of the best estimate values.
The empirical arrival time data used for the likelihood analysis are necessarily an overestimation of the actual date of the importation of cases as cases could go undetected. If we assume a shift of 7 days earlier for all arrival times available from official reports, the resulting maximum likelihood is increasing the best estimate for R_0 to 1.87 (95% CI 1.73 to 2.01), as expected since earlier case importation necessitates a larger growth rate of the epidemic. The official timeline used here therefore provides, all other parameters being equal, a lower estimate of the transmission potential. We have also explored the use of a subset of the 12 countries, always generating results within the confidence interval of the best estimate.
The best estimates reported in Table (7.5) do not show any observable dependence on the assumption about the seasonality scenario. The analysis is restricted to the first countries seeded from Mexico to preserve the conditional independence of the variables and it is natural to see the lack of any seasonal signature since these countries receive the disease from a single country , mostly found in the tropical region where no seasonal effects are expected.
In order to find the minimum seasonality scaling factor α_{min} that best fits the empirical data, we performed a statistical correlation analysis of the arrival time of the infection in the 93 countries infected as of 18 June. By considering a larger number of countries and a longer period for the unfolding of the epidemic worldwide as seasons change, the correlation analysis for the baseline scenario provides clear statistical indications for a minimum rescaling factor in the interval $0.6 < \alpha_{min} < 0.7$. In the full range of epidemic parameters explored, the correlation analysis yields values for α_{min} in the range 0.4 to 0.9. This evidence for a mild seasonality rescaling is consistent with the activity observed in the months of June and July in Europe and the US where the epidemic progression has not stopped and the number of cases keeps increasing considerably (see also Table (7.6) for the corresponding values of $R(t)$ in those regions during summer months).
This analysis allows us to provide a comparison with the epidemic activity observed and an early assessment of the future unfolding of the epidemics. For each set of parameters the model generates quantities of interest such as the profile of the epidemic behavior in each subpopulation or the number of imported cases. Each simulation generates a stochastic realization of the process and the curves are the statistical aggregate of at least

$2 \cdot 10^3$ realizations. In the following we report the median profiles and where indicated the 95% CI. Results are in good agreement with the reported temporal evolution of the epidemic and highlight a progressive decrease of the monitoring activity caused by the increasing number of cases, as expected (158). The same information is also available for each single subpopulation defined in the model. We have therefore tested the model results in four territories of Australia. Interestingly, the model is able to recover the different timing observed in the four territories. In Figure (7.6)-c-d we report the predicted baseline case profiles for countries in the Southern hemisphere. It is possible to observe in the figure that in this case, the effect of seasonality is not discriminating between different waves, as the short time interval from the start of the outbreak to the winter season in the Southern hemisphere does not allow a large variation in the rescaling of the transmissibility during these months. Therefore we predict a first wave that occurs between August and September in phase with the seasonal influenza pattern, and independently of the seasonality parameter α_{min}. The situation is expected to be different in the Northern hemisphere where different seasonality parameters might progressively shift the peak of the epidemic activity in the winter months. Figure (7.6)-e reports the predicted daily incidence profiles for the Northern hemisphere and the 95% CI for the activity peaks of the pandemic with the best-fit seasonality scenario (that is, the range $0.6 < \alpha_{min} < 0.7$). Table (7.7) reports the same information for different continental areas. The general evidence clearly points to the occurrence of an autumn/winter wave in the Northern hemisphere strikingly earlier than expected, with peak times ranging from early October to the middle of November. The peak estimate for each geographical area is obtained from the epidemic profile summing up all subpopulations belonging to the region. The activity peak estimate for each single country can be noticeably different from the overall estimate of the corresponding geographical region as more populated areas may dominate the estimate for a given area. For instance Chile has a pandemic activity peak in the interval 1 July - 6 August, one month earlier than the average peak estimate for the Lower South America geographical area it belongs to. It is extremely important to remark that in the whole phase space of parameters explored the peak time for the epidemic activity in the Northern hemisphere lies in the range late September to late November, thus suggesting that the early seasonal peak is a genuine feature induced by the epidemic data available.

In Table (7.8) we report the new number of cases at the activity peak and the epidemic

size as of 15 October for a selected number of countries. As shown by the results in the table, in Figure (7.7), and in more details in another our paper (110), the massive vaccination campaign was negligible as it was expected in case of an early peak scenario predicted by our simulations, because most of the vaccine doses were not deployed before November 2009 (see Table (7.4) for details). The timing of the epidemic activity is a key aspect in the public health decision making. Our results were submitted for publication in late July 2009. How good were they? This is an important questions to address, that define the accurancy and validity of our model. We compared the simulated epidemic timeline with the epidemiological data reported by the surveillance systems of several countries worldwide. In Figure (7.7) shows the peak times of the simulated pandemic runs, on a weekly basis, and the incidence peak weeks reported by the national surveillance systems during the $2009-2010$ Winter season, for 46 countries in the Northern Hemisphere. To take into account the uncertainty related to the surveillance reporting systems, we displayed the observed peak weeks as a color gradient, whose limits correspond to the time interval where an incidence higher than 80% of the maximum incidence was observed. In the Northern Hemisphere, most of the countries experienced a single major pandemic wave during autumn. The predominant strain of the $2009-2010$ winter season in all countries was the 2009 A(H1N1) pandemic strain, accounting for more than 90% of all the influenza virus specimens analyzed worldwide (194). The influenza activity peaked, during the October - December period, that is much earlier than the usual timing of seasonal influenza, generally peaking between January and March. The epidemic wave peaked first in North America, at the end of October, and later in Europe with a wide range of peak weeks, from late October in Iceland to late December in Serbia. The observed peak weeks are in good agreement with the model results, as they all lie within the 95% reference range of the simulations.

In order to assess the amount of pressure on the healthcare infrastructure, in Table (7.9) we provide the expected number of hospitalizations at the epidemic peak according to different hospitalization rate estimates. The assessment of the hospitalization rate is very difficult as it depends on the ratio between the number of hospitalizations and the actual number of infected people. As discussed previously, the number of confirmed cases released by official agencies is always a crude underestimate of the actual number of infected people. We consider three different methods along the lines of those developed for the analysis of fatalities due to the new virus (195). The first assumes the average value

Country	Mass vaccination starting date	Final population coverage
China	September 14^{th}, 2009	6%
Hungary	October 1^{st}, 2009	30%
United States	October 5^{th}, 2009	27%
Canada	October 12^{th}, 2009	45%
Italy	October 12^{th}, 2009	1.5%
Japan	October 19^{th}, 2009	17%
Israel	October 19^{th}, 2009	9%
France	October 20^{th}, 2009	9%
Sweden	October 21^{st}, 2009	60%
United Kingdom	October 26^{th}, 2009	8%
Germany	October 26^{th}, 2009	8%
Portugal	October 26^{th}, 2009	3%
Finland	October 26^{th}, 2009	50%
Austria	October 26^{th}, 2009	3.3%
Ireland	October 31^{st}, 2009	17%
Denmark	November 2nd, 2009	6%
Turkey	November 2nd, 2009	3%
Iceland	November 2nd, 2009	40%
Belgium	November 2nd, 2009	7.5%
Slovenia	November 2nd, 2009	5%
Netherlands	November 9^{th}, 2009	25%
Switzerland	November 15^{th}, 2009	15%
Spain	November 16^{th}, 2009	4.5%
Greece	November 16^{th}, 2009	3%
Tunisia	November 16^{th}, 2009	2.6%
Czech Republic	November 23^{rd}, 2009	0.6%
Norway	December 1^{st}, 2009	45%

Table 7.4: Mass vaccinations in Northern Hemisphere countries during the 2009 - 2010 Winter season.

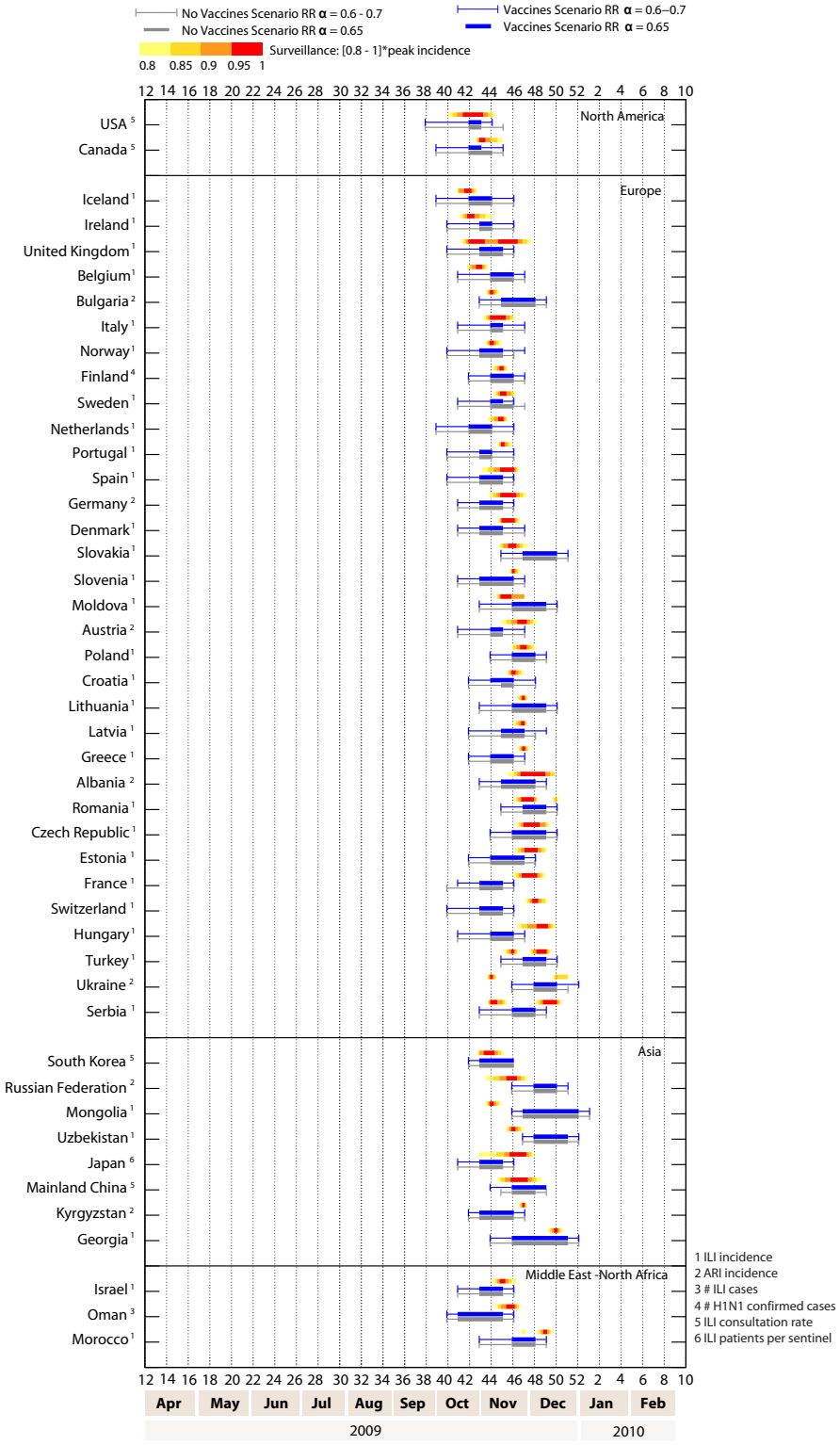

Figure 7.7: Comparison between real data and simulations with and without vaccinations and different values of α_{min} in 46 countries.

of hospitalization observed during the regular seasonal influenza season. The second is a multiplier method in which the hospitalization rate is obtained as the ratio between the WHO number of confirmed hospitalizations and the cases confirmed by the WHO multiplied by a factor 10 to 30 to account for underreporting. The third method is given by the ratio of the total number of confirmed hospitalizations and the total number of confirmed cases. This number is surely a gross overestimation of the hospitalization rate (195; 196). It has to be noted that hospitalizations are often related to existing health conditions, age and other risk factors. This implies that hospitalizations will likely not affect the population homogenously, a factor that we cannot consider in our model.
The number of hospitalized at peak times in the selected countries range between 2 and 40 per 100.000 persons, for a hospitalization rate typical of seasonal influenza and for an assumed 1% rate, respectively, yielding a quantitative indication of the potential burden that the health care systems will likely face at the peak of the epidemic activity in the next few months. It is worth noting that the present analysis considers a worst-case scenario in which no effective containment measures are introduced. This is surely not the case in that pandemic plans and mitigation strategies are considered at the national and international level. Guidelines aimed at increasing social distancing and the isolation of cases will be crucial in trying to mitigate and delay the spread in the community, thus reducing the overwhelming requests on the hospital systems. Most importantly, the mass vaccination of a large fraction of the population would strongly alter the presented picture. By contrast, any mass vaccination started before the middle of October as shown in Table (7.4). As we said the early activity peak of the pandemic in October/November reduced a lot the effectiveness of vaccination program that took place too late with respect to the pandemic wave in the Northern hemisphere. In this case it is natural to imagine the use of other mitigation strategies aimed at delaying the activity peak so that the maximum benefit can be gained with the vaccination program. As an example, we studied the implementation of systematic antiviral (AV) treatment and its effect in delaying the activity peak (145; 160; 162; 169; 197; 198; 199; 200). The resulting effects are clearly country specific in that each country will experience a different timing for the epidemic peak (with a local transmissibility increasing in value as we approach the winter months) and will count on antiviral stockpiles of different sizes. Here we consider the implementation of the AV treatment in all countries in the world that have drugs stockpiles available (source data from (201; 202) and national agencies),

Parameter	Best Estimate	95% CI	Description
R_0	1.75	1.64 to 1.88	Basic reproduction number
G_t	3.6	2.2 to 5.1	Mean generation time (days)
μ^{-1}	2.5	1.1 to 4.0	Mean infectious period (days)
α_{min}	0.65	0.6 to 0.7	Winter minimal seasonality rescaling

Assumed values	Best Estimate	Sensitivity analysis range	Description
ϵ^{-1}	1.1	1.1 to 2.5	Mean exposed period (days)
α_{max}	1.1	1.0 to 1.1	Summer maximum seasonality rescaling

Table 7.5: Best Estimates of the epidemiological parameters. Estimates from the Monte Carlo likelihood analyses for various values of the parameter space explored. The confidence interval is determined by the likelihood procedure.

until the exhaustion of their stockpiles (151). We have modeled this mitigation policy with a conservative therapeutic successful use of drugs for 30% of symptomatic infectious individuals. The efficacy of the AV is accounted in the model by a 62% reduction in the transmissibility of the disease of an infected person under AV treatment when AV drugs are administered in a timely fashion (160; 162). We assume that the drugs are administered within 1 day of the onset of symptoms. We also consider that the AV treatment reduces the infectious period by 1 day (160; 162). In Figure (7.8) we show the delay obtained with the implementation of the AV treatment protocol in a subset of countries with available stockpiles. As an example, we also show the incidence profiles for the cases of Spain and Germany, where it is possible to achieve a delay of about 4 weeks with the use of 5 million and 10 million courses of AV, respectively. The results of this mitigation might be extremely valuable in providing the necessary time for the implementation of the mass vaccination program.

7.2.4 Estimating the initial number of cases in Mexico

By using GLEaM it is possible to provide a model estimate of the number of imported cases arriving from Mexico to a set of selected countries. The estimated 99% reference

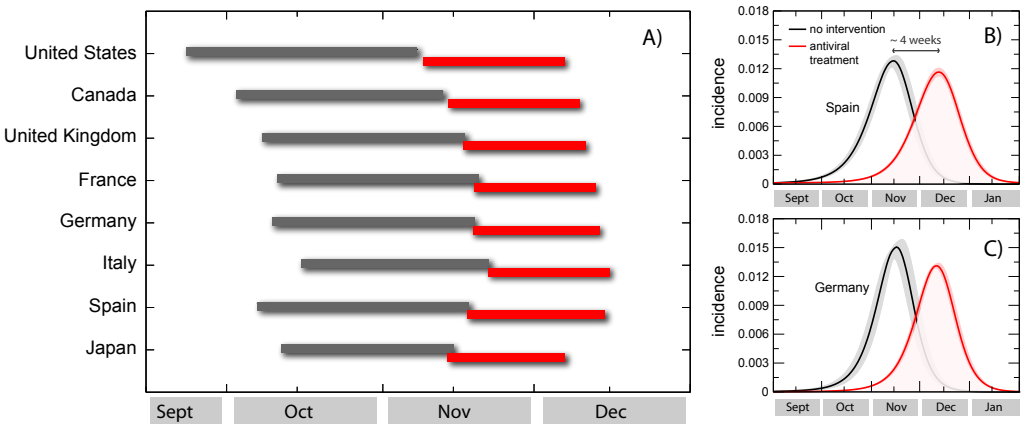

Figure 7.8: Delay effect induced by the use of antiviral drugs for treatment with 30% case detection and drug administration. (a) Peak times of the epidemic activity in the worst-case scenario (black) and in the scenario where antiviral treatment is considered (red), for a set of countries in the Northern hemisphere. The intervals correspond to the 95% confidence interval (CI)of the peak time for the two values defining the best-fit seasonality scaling factor interval. (b,c) Incidence profiles for Spain and Germany in the worst-case scenario (black) and in the scenario where antiviral treatment is considered (red). Results are shown for $\alpha_{min} = 0.6$ only, for the sake of visualization. A delay of about 4 weeks results from the implemented mitigation

Month	$R(t)$ in Northern hemisphere
May	1.19 to 1.49
June	1.07 to 1.33
July	1.05 to 1.24
August	1.07 to 1.33
September	1.19 to 1.49

Table 7.6: Seasonality time-dependent reproduction number in the Northern hemisphere. The values of $R(t)$ for the Northern hemisphere correspond to the rescaling of the maximum likelihood value of R_0 in Mexico and in the Tropical regions ($R_0 = 1.75$) and the best values for the seasonality rescaling factor, $0.6 < \alpha_{min} < 0.7$. The parameter α_{min} indicates the minimum value of the seasonal rescaling of R_0 induced by the sinusoidal forcing in the Northern hemisphere (154).

Region	Estimated activity peak time
North America	25 September to 9 November
Western Europe	14 October to 21 November
Lower South America	30 July to 6 September
South Pacific	28 July to 17 September

Table 7.7: Peak times. The table reports the 95% confidence interval (CI) for the pandemic activity peak time for geographical areas in the Northern and Southern hemispheres estimated for the best-fit seasonality scaling interval, $0.6 < \alpha_{min} < 0.7$, and for the maximum likelihood value of R_0 found for the baseline parameters, $R_0 = 1.75$. The confidence interval is obtained from the set of numerical observations of the peak time in a given region obtained from the $2 \cdot 10^3$ stochastic runs of the model. In all cases we obtain activity peak time intervals close to those reported for the baseline scenario.

		New daily cases at the peak time	New daily cases at the peak time	Epidemic Size at 15 October (% of population)	
Country	Peak Time	(thousands)	(% of population)	$\alpha_{min}0.6$	$\alpha_{min}0.7$
United States	24 Sep to 9 Nov	2,983 to 3,302	1.06 to 1.17	4.99 to 7.38	23.76 to 29.96
Canada	4 Oct to 14 Nov	331 to 373	1.04 to 1.17	2.28 to 4.56	16.90 to 27.41
United Kingdom	9 Oct to 18 Nov	723 to 813	1.21 to 1.36	1.77 to 4.45	11.11 to 27.29
France	12 Oct to 21 Nov	725 to 792	1.26 to 1.38	1.83 to 3.87	10.86 to 26.40
Germany	11 Oct to 20 Nov	1,162 to 1,291	1.43 to 1.59	1.02 to 2.41	8.57 to 26.25
Italy	17 Oct to 23 Nov	793 to 867	1.39 to 1.52	0.93 to 2.20	6.71 to 22.13
Spain	8 Oct to 19 Nov	492 to 536	1.23 to 1.34	2.39 to 3.70	13.26 to 27.95
China	8 Nov to 11 Dec	14,077 to 16,207	1.16 to 1.34	0.65 to 5.34	1.51 to 9.49
Japan	13 Oct to 16 Nov	1,539 to 1,822	1.21 to 1.43	1.47 to 4.86	5.84 to 24.65

Table 7.8: Peak times of the epidemic activity, daily new number of cases predicted at peak time and % of the population, and epidemic size on 15 October are shown. Intervals refer to the 95% confidence interval (CI). After 1 year from the start of the epidemic the percentage of total population infected is close to 45% with small differences of the order of 5% across different countries.

	Seasonal influenza	Multiplier method		WHO confirmed cases
	HR:0.08%	HR:0.3%	HR:1%	HR:10%
USA	2.21	8.28	27.58	275.84
Canada	2.18	8.17	27.22	272.23
UK	2.52	9.45	31.52	315.15
France	2.61	9.79	32.64	326.40
Germany	2.98	11.17	37.22	372.18
Italy	72.87	10.76	35.87	358.67
Spain	2.54	9.54	31.81	318.12
China	2.48	9.32	31.05	310.50
Japan	2.59	9.70	32.32	323.19

Table 7.9: Number of hospitalizations per 100,000 persons at the activity peak in several countries. The estimates are obtained by considering three methods. The first assumes the average hospitalization rate (HR) observed during the seasonal influenza season. The second is a simple multiplier method in which the HR is obtained as the ratio between the World Health organization (WHO) number of confirmed hospitalizations and the cases confirmed by the WHO multiplied by a factor 10 to 30 to account for underreporting. The third method is simply the ratio of the total number of confirmed hospitalizations and the total number of confirmed cases.

range is shown in Table (7.10). The dates and target countries are chosen to facilitate the comparison with the numbers found in the literature (203; 204; 205; 206). The numbers shown in the Table refer to the importation of infected/exposed individual traveling from Mexico in one of the listed countries as of the date of May the 8^{th}. Only 2/3 of the exposed travelers are then considered in the cumulative number of cases as only this fraction will eventually develop symptoms, according to the model assumptions. The numbers of imported cases to each country are typically small, and as such prone to large stochastic fluctuations. However the surveillance values are all within the 99% reference ranges of the $2 \cdot 10^3$ realizations of our model. We provided in Ref. (16) a full sensitivity analysis of the results but we observe very small variations with respect to the presented results in the range of parameters explored. This is because any Maximum Likelihood Estimate (MLE) for R_0 and generation interval tend to optimize the growth rate with respect to the epidemic timeline thus producing very similar results in the early spreading of the epidemic. We have also considered that in the US the travel history is known only for 50% of the confirmed cases. The simple extrapolation that provides a twofold estimate of imported cases (in brackets in Table (7.10)) is however still compatible with the reference range of our stochastic simulations.

Table (7.11) shows GLEaM predictions for the size of the epidemic in Mexico on April 30^{th} and compare the results with the estimations of Refs. (157) and (203). We provide the 95% reference range over 210^3 realizations. The obtained range includes the lower bound estimate of Ref. (203). Our median value for the number of asymptomatic cases is 734.000 that is again compatible with the range of values reported in Ref.(203). While the estimates presented in Refs. (157) and (203) are based on a homogeneous mixing approach within the entire Mexico, the approach used here is a spatially structured model that just in Mexico counts 65 different census areas. These census areas are not equally connected internationally and between them. The number of cases relevant for the international spread of infected individuals are mostly in census areas close to international transportation hubs. Poorly connected regions of Mexico on the other hand, while experiencing a considerable number of cases, would contribute only marginally to the International spread of cases. This observation readily explains why single population calculations that match the detection of imported cases with the local prevalence are necessarily underestimating the latter quantity.

Number imported cases (May 8th)	USA	UK	France	Germany	Brazil
Simulation Results	0 - 534	0 - 44	0 - 62	0 - 55	0 - 45
Surveillance data	85 (170)	17	11	9	3

Table 7.10: Cumulative number of imported cases from Mexico shown as the 99% reference range over $2{\cdot}10^3$ realizations on May 8 for a few countries. The simulations are obtained with the best estimate parameters of the baseline case of Ref. (16) and R_0=1.75 [95%CI 1.64 to 1.88]. The number of imported infected individuals and of independent clusters correspond to the data given in Ref. (203) for US, and UK and the values in (205) for France, in (204) for Germany and in (206) for Brazil. No data was available to assess the possible presence of clusters in Germany and France. In the USA we report in parentheses the revised number considering the rate of unknown travel history in confirmed cases.

While GLEaM takes into account a higher level of geographical organization than previous approaches, its estimates still contain a number of assumptions and approximations. The contagion within each census area is approximated by means of a homogeneous mixing process. Once a person arrives at a census area by plane, he/she comes integrated into the local population. This implies that, as in (203), the travelers and the local population are equally exposed to the disease. Finally, the model considers each individual as independent and the possibility of cluster cases is not considered. Despite these shortcomings and other necessary uncertainties, GLEaM predictions might provide additional information for a better understanding of the early evolution of the present pandemic. Despite the different approximations used here and in Ref.(203), both approaches are providing support to the possibility of a reporting ratio of infected cases in Mexico as low as 1 in 100, in agreement with prior estimates (113). This finding is important when evaluating the massive amount of data which are now being collected in a large number of countries around the world. We can easily imagine that the reporting rate as well as any estimate of the cumulative attack rate in most of the countries could be easily underestimated by orders of magnitude.

7.2.5 Modeling the critical care demand and antibiotics resources

We model the administration of vaccines through a dynamic vaccination campaign with a uniform daily rate r_v of distribution to the population in countries where doses are

	Number of symptomatic cases in Mexico (Apr. the 30th)
Simulation Results	[121,000 - 1,394,000]
Lower bound range of Ref. (203)	113,000-375,000
Estimate of Ref. (157)	2,000 - 280,000
Mexican official report (207) (confirmed cases)	3,350

Table 7.11: Predictions of GLEaM for the size of the epidemic in Mexico on April 30 in thousands of cases and comparison with other approaches and with empirical data. The simulations are obtained with the best estimate parameters of the baseline case of Ref. (16) and show the 95% reference range over 210^3 stochastic realizations. The results are compared with the lower bound estimate range in (203), the estimate provided in Ref. (157) and the number of confirmed cases given by official reports. The interval provided for Ref.(157) is obtained by merging the results reported in the paper under different assumptions and including the 95% CI.

available, till their exhaustion. We explore two values for the daily distributions rate, $r_v = 0.1\%$ consistent with the current availability of doses and distribution in several countries, and $r_v = 1\%$ based on the distribution policies planned during Summer 2009 on the timing of vaccine development and testing (110). We assume the administration of a single dose of vaccine, providing protection with a delay of 2 weeks. A full description of the vaccination implementation and sensitivity analysis is reported in Ref. (110). Following the estimates of the severity of H1N1 pandemic, we assume a complication rate of 15% of clinical cases (208), a hospitalization rate of 0.5% of clinical cases (209), and an ICU admission rate of 15% of hospitalized patients (210). We model influenza-related pneumonia as a complication associated to influenza infection, considering two main types of pneumonia: primary viral pneumonia and secondary bacterial pneumonia. While bacterial coinfection was shown to be the predominant cause of death in previous influenza pandemics (211), its presence in the severe cases analyzed since the start of the outbreak range from almost no evidence in the early reviews (212; 213; 214), to about 10% (215), 33% or larger proportions (216; 217) of the cases presenting influenza-associated complications. These fluctuations in the role of bacterial pneumonia might be due to the difficulty of testing for specific bacterial diagnosis, or to the use of antibiotics prior to routine clinical tests. Given the uncertainty on the cause of pneumonia

at this stage of the epidemic evolution, we assume a proportion of bacterial pneumonia in cases showing complications in the range of $\alpha = 33 - 50\%$, with a sensitivity exploring a 10% proportion. Under pandemic conditions, it is assumed that very small differences will be implemented in the management and treatment of the patients with either types of pneumonia, as the diagnosis of influenza-associated complications will be mostly based on clinical findings and most prescribing will be empirical, based on both antibacterial therapy and antiviral medications (218). Multiple subsequent stages of pneumonia course are modeled according to the CURB-65 classification score (219) as reported in Table (7.12), and different progressions are assumed to take into account both viral and bacterial pneumonia (see Figure (7.9)). It is also worth remarking that the model does not consider social structure in the subpopulations, therefore the effect of prioritized distribution of vaccines to individuals belonging to risk groups in reducing the number of hospitalizations and deaths is not considered here. These assumptions represent a necessary trade-off for the computational efficiency of the model that allows to perform parameter estimations fitting the worldwide pattern of the pandemic (16), explore several scenarios under different conditions, and perform sensitivity analysis on the assumptions. Once the disease parameters and initial conditions are defined, GLEaM generates in-silico epidemics for which we can gather information such as incidence and prevalence of all stages considered in the compartmentalization, for each subpopulation in the world and with a time resolution of one day. All results shown in the following sections are obtained from the statistics based on at least $2 \cdot 10^3$ stochastic runs of the model.

Based on the available knowledge of complication, hospitalization and ICU rates, and the relative proportion of bacterial vs. viral pneumonia, the simulation results allow the measure of the predicted need of beds in intensive care units, and provide estimates of the corresponding courses of antibiotics needed. Figure (7.10) shows the time evolution of the predicted prevalence of ICU occupancy for a given set of countries. In the baseline case, when no intervention is implemented, the ICU prevalence peak ranges between approximately 5 and 7 ICU beds per 10^5 people. These values are well below the national average capacity of some countries, such as e.g. the United States with a total of about 20 ICU beds per 10^5 (220) and Germany with an average of approximately 28 ICU beds per 10^5 (221). The predicted need is slightly lowered if a 0.1% dynamic vaccination is considered, and would be reduced to values in the range of 3.6 to 4.8 ICU

beds per 100,000 if we assume $r_v = 1\%$, below the national average number of ICU beds of many European countries (222). While the predicted ICU beds needs are averaged at the country level to conform with the capacity data, it is however important to note that the impact and the potential occurrence of critical situations strongly depends on the geographic distribution of the critical care resources, with areas that might have access to a larger number of intensive care units than others (see for example Ref. (223)). Moreover, a direct comparison between the simulated demand and critical care availability is made difficult by the lack of a standard definition for intensive care unit beds, and the large variations observed in both numbers of beds and volume of admission between countries in North America and Western Europe (222).

The results shown in Figure (7.10) are based on an average ICU length of staying equal to $L_{ICU} = 7$ days. Since there is a large variation in this parameter, with cohort studies showing median duration of 7 days and interquartile range up to approximately 2 weeks (217), we also explored the effect of considering longer lengths of staying, $L_{ICU} = 10$ and $L_{ICU} = 14$ days. The longer bed occupancy would inevitably lead to an increase in the need of ICU beds at peak, in the range of approximately 9 to 12 per 100,000 persons in the case of 14 days of average ICU duration (see Table (7.12)).

Table (7.14) reports the number of antibiotics courses needed daily at the peak of the requests, and the total size predicted to be used at the end of the pandemic wave, based on the empirical guidelines of the British Thoracic Society (219; 224) and broken down by the stage of severity of pneumonia. A single course of antibiotics is defined as the combination of antimicrobial drugs considered in the treatment regimen for the suggested duration (see Table (7.12)). In the case of non severe pneumonia, the predicted need for antibiotics at peak usage is in the range of $[150 - 230]$ courses per 100,000 with variations depending on the country under study, under the assumption that no intervention is considered. The total size of antibiotics courses predicted to be used in the current Fall 2009 pandemic is in the range of $[6,337 - 7,149]$ per 100,000, which needs to be compared with the available stockpiles of antibiotics courses to cover high-risk groups. Many countries however do not possess nation-wide antibiotic supplies, and the estimates contained in Table (7.14) can therefore be considered as guidelines to assess the expected needs during the remaining evolution of the pandemic wave with respect to the present usage pattern and available resources.

Along with anecdotal reports indicating ICUs being overwhelmed by the sudden surge

of H1N1 cases with severe complications (224), studies on the Winter experience in the Southern Hemisphere during the H1N1 pandemic wave confirm a substantial impact on ICUs, with the maximum number of ICU beds occupied by region in Australia and New Zealand ranging between 0.63 and 1.1 per 100,000 inhabitants (217). These values are smaller than the ICU demands predicted for the Fall wave in the Northern Hemisphere. It is important to note, however, that the used model does not take into account the population structure (age dependent attack rates), risk groups and prior immunity thus likely overestimating the global attack rate of the pandemic. Furthermore we do not include in the model mitigation factors (e.g. social distancing, targeted school closures, etc.) that might have contributed to the reduction of the overall burden on the critical care facilities in the Southern Hemisphere; a similar reduction on burden could also be seen in the Northern Hemisphere.

Severity of complications	Assessment	Recommended action/compart.	Average duration
non-severe pneumonia	CURB-65=0-2	home treatment or supervised outpatient treatmen	3.5 days (212)
severe pneumonia	CURB-65=3 or presence of bilateral lung infiltrates on chest x ray	hospital ward	1.5 days to ICU admission (hospital ward 1), 5 days to recovery (hospital ward 2) (215)
	CURB-65=4-5 or bilateral chest x ray changes	ICU	7, 10, 14 days (215; 217)

Table 7.12: Severity assessment, recommended action, and estimated durations assumed in the model. We refer to CURB-65 score as the method used to determine the management of influenza-related complications in patients admitted to hospital (219). CURB-65 score is calculated by assigning one point for each of the following: Confusion (mental test score of ≤ 8, or new disorientation in person, place or time),Urea > 7 $mmol/l$, Respiratory rate $\geq 30/min$, Blood pressure (SBP$<$ 90 mmHg or DBP$\leq$ 60 mmHg), Age ≥ 65 years. Three subsequent stages are defined to model complications, based on the recommended action. Patients with bilateral lung infiltrates on chest radiography consistent with viral pneumonia are assumed to be managed as severe pneumonia, regardless of CURB-65 score (219). The preferred empirical antibiotic regimens for treatment of patients in each stage are based on the guidelines issued by the British Thoracic Society (219; 224). Patients in home treatment and hospital ward are assumed to take co-amoxiclav 625 mg tds PO or doxycycline 200 mg stat and 100 mg od PO for 7 days, and patients in ICU are assumed to take co-amoxiclav 1.2 g tds IV or cefuroxime 1.5 g tds IV or cefotaxime 1g tds IV plus Macrolide (erythromycin 500 mg qds IV or clarithromycin 500 mg bd IV) for 10 days. All patients at all stages of severity of complications are also expected to receive antivirals, with a dosage of 2 tablets per day.

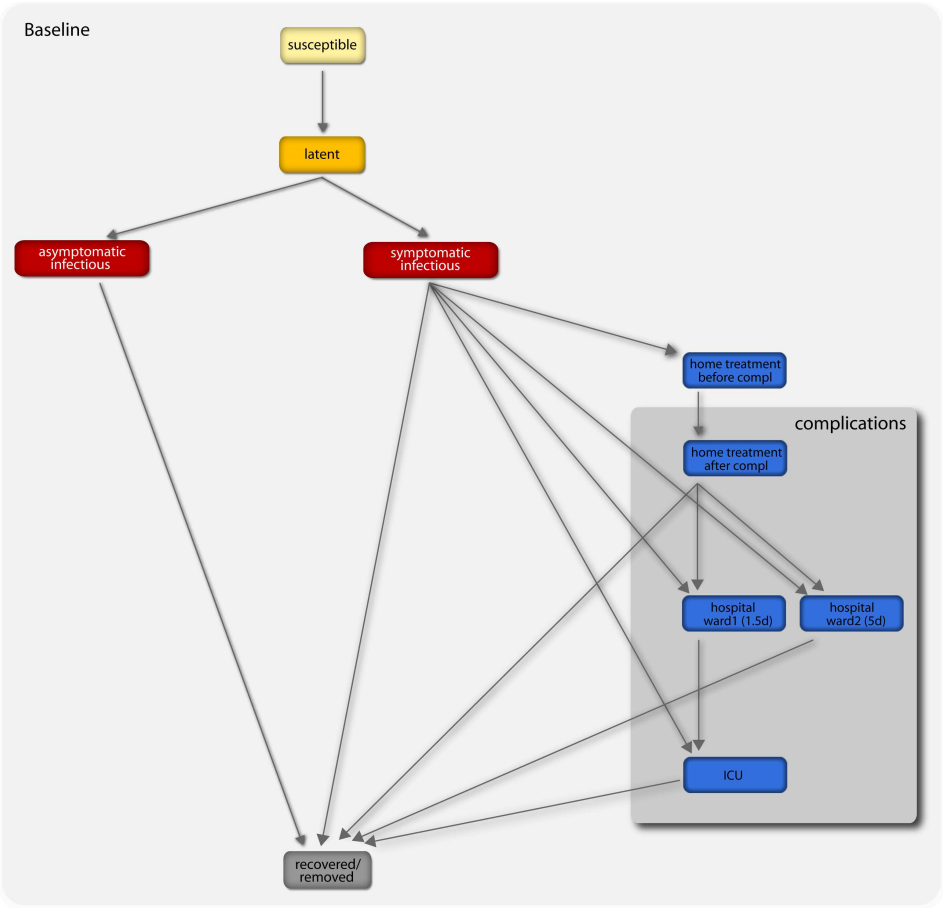

Figure 7.9: Diagram flow of the transmission model. A susceptible individual interacting with an infectious person may contract the illness and enter the latent compartment where he is infected but not yet infectious. At the end of the latency period, each latent individual becomes infectious entering the symptomatic compartment with probability $(1 - p_a)$ or becoming asymptomatic with probability p_a. Asymptomatic individuals infect with a reduced transmission rate. A fraction $(1 - p_t)$ of the symptomatic individuals would stop traveling when ill. A full description of the parameter values is reported in Ref. (21). If vaccines are available, a fraction equal to r_v of the susceptible population enters the susceptible vaccinated compartment each day. A similar progression to the baseline compartmentalization is considered if infection occurs (see Ref. (110)). The model assumes that infectious individuals might develop complications with a rapid progression to severe conditions requiring hospitalization or ICU admission (i.e. second and third stage of the complications tree, respectively), or home treatment (i.e. first stage) with pneumonia symptoms appearing during the early convalescent period of the influenza infection (219). The compartments 'hospital ward 1' and 'hospital ward 2' refer to different lengths of staying of the patient in the hospital ward (see Table (7.12)), depending on subsequent worsening of symptoms or direct recovery, respectively. Progressions from one stage to the others is modeled according to the average length of staying in each compartment as obtained from clinical studies (215; 217) (see also Table (7.12)) and based on the available estimates of complication, hospitalization and ICU admission rates (208; 209; 210).

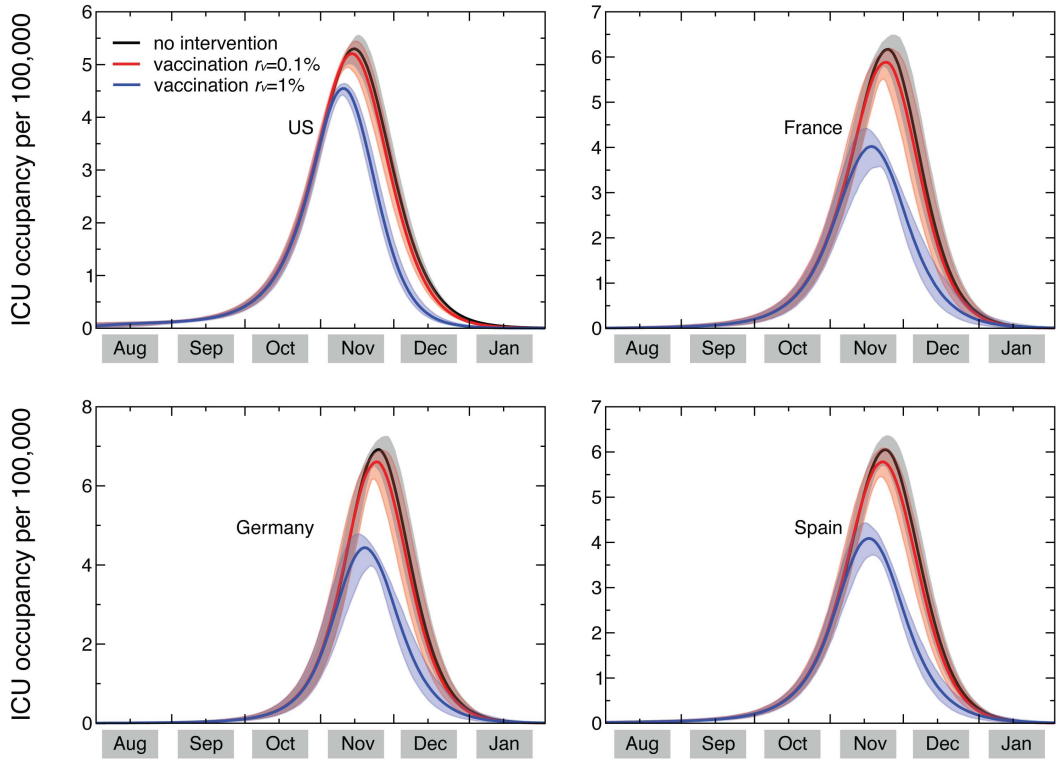

Figure 7.10: Time evolution of the ICU occupancy in a set of countries. ICU occupancy measures the predicted need of ICU beds per 100,000 persons. Results for the United States, France, Germany, and Spain are shown. The three profiles per each country refer to the predicted ICU occupancy in the baseline case when no intervention is implemented, and in case dynamic vaccination campaigns with distribution rates $r_v = 0.1\%$ and $r_v = 1\%$ are considered. Solid curves correspond to the median profiles and the shaded areas to the 95% reference range obtained from 2,000 stochastic simulations. The average ICU length of staying is assumed equal to 7 days (217).

ICU occupancy at peak (per 100,000)									
Country	**Baseline**			**Vaccination campaigns**					
				0.1%			1%		
	7 days	10 days	14 days	7 days	10 days	14 days	7 days	10 days	14 days
US	[5.0-5.6]	[6.8-7.5]	[8.7-9.7]	5.0-5.5]	[6.7-7.3]	[8.6-9.4]	[4.5-4.6]	[5.9-6.2]	[7.6-7.9]
UK	[5.7-6.5]	[7.6-8.6]	[9.9-11.0]	[5.5-6.2]	[7.4-8.2]	[9.6-10.5]	[3.9-4.6]	[5.2-6.1]	[6.7-7.7]
Canada	[5.0-5.7]	[6.7-7.6]	[8.7-9.9]	[4.8-5.5]	[6.5-7.3]	[8.5-9.5]	[3.8-4.4]	[5.1-5.8]	[6.5-7.3]
France	[5.9-6.6]	[7.9-8.7]	[10.2-11.2]	[5.7-6.2]	[7.6-8.3]	[9.8-10.6]	[3.6-4.4]	[4.9-5.9]	[6.3-7.4]
Italy	[6.5-7.1]	[8.6-9.4]	[11.0-12.0]	[6.2-6.7]	[8.2-8.9]	[10.5-11.3]	[3.6-4.5]	[4.8-5.9]	[6.1-7.4]
Spain	[5.8-6.4]	[7.8-8.6]	[10.0-11.0]	[5.6-6.1]	[7.5-8.2]	[9.6-10.5]	[3.8-4.5]	[5.1-5.9]	[6.5-7.5]
Germany	[6.6-7.3]	[8.8-9.7]	[11.2-12.2]	[6.4-7.0]	[8.5-9.2]	[10.8-11.6]	[4.0-4.8]	[5.4-6.4]	[6.8-8.0]

Table 7.13: Predicted need of ICU beds in the baseline case scenario and in the case of vaccination campaigns. The 95% reference range (RR) of the daily number of occupied ICU beds per 100,000 is reported at its peak for several countries in the Northern Hemisphere.

Antibiotic usage – baseline						
Country	**Daily administered AB courses at peak (per 100,000)**			**Total administered AB courses at the end of pandemic wave (per 100,000)**		
	Pneumonia stage I	**Pneumonia stage II**	**Pneumonia stage III**	**Pneumonia stage I**	**Pneumonia stage II**	**Pneumonia stage III**
US	[152-171]	[4.4-4.9]	[0.8-0.9]	[6,196-6,455]	[183-191]	[31.7-33.0]
UK	[176-197]	[5.1-5.8]	[0.9-1.1]	[6,529-6,845]	[193-203]	[33.3-35.1]
Canada	[150-170]	[4.4-5.0]	[0.8-1.0]	[6,508-6,755]	[192-200]	[33.0-34.8]
France	[184-201]	[5.3-5.9]	[1.0-1.1]	[6,611-6,906]	[195-204]	[33.7-35.4]
Italy	[202-221]	[5.8-6.4]	[1.1-1.2]	[6,758-6,981]	[200-206]	[34.4-35.8]
Spain	[178-195]	[5.2-5.7]	[0.9-1.1]	[6,584-6,815]	[194-202]	[33.4-35.1]
Germany	[208-230]	[5.9-6.6]	[1.1-1.2]	[6,739-6,990]	[199-207]	[34.4-35.8]
Antibiotic usage – vaccination with $r_v = 0.1\%$						
	Daily administered AB courses at peak (per 100,000)			**Total administered AB courses at the end of pandemic wave (per 100,000)**		
	Pneumonia stage I	**Pneumonia stage II**	**Pneumonia stage III**	**Pneumonia stage I**	**Pneumonia stage II**	**Pneumonia stage III**
US	[151-166]	[4.4-4.8]	[0.8-0.9]	[6,005-6,220]	[177-184]	[30.7-31.9]
UK	[170-186]	[4.9-5.4]	[0.9-1.0]	[6,297-6,540]	[186-193]	[32.1-33.6]
Canada	[147-164]	[4.3-4.9]	[0.8-0.9]	[6,278-6,457]	[185-191]	[31.8-33.3]
France	[176-188]	[5.1-5.5]	[0.9-1.0]	[6,357-6,585]	[188-195]	[32.3-33.8]
Italy	[191-206]	[5.5-6.0]	[1.0-1.1]	[6,481-6,633]	[191-196]	[32.9-34.1]
Spain	[171-185]	[5.0-5.4]	[0.9-1.0]	[6,335-6,511]	[187-193]	[32.1-33.6]
Germany	[200-216]	[5.7-6.2]	[1.0-1.2]	[6,476-6,654]	[191-197]	[33.0-34.2]
Antibiotic usage – vaccination with $r_v = 1\%$						
	Daily administered AB courses at peak (per 100,000)			**Total administered AB courses at the end of pandemic wave (per 100,000)**		
	Pneumonia stage I	**Pneumonia stage II**	**Pneumonia stage III**	**Pneumonia stage I**	**Pneumonia stage II**	**Pneumonia stage III**
US	[140-144]	[4.0-4.1]	[0.7-0.8]	[4,801-4,862]	[142-144]	[24.5-25.0]
UK	[120-140]	[3.5-4.1]	[0.6-0.8]	[4,452-4,762]	[131-141]	[22.7-24.5]
Canada	[121-133]	[3.5-3.9]	[0.6-0.8]	[4,517-4,732]	[133-140]	[22.9-24.4]
France	[110-136]	[3.2-4.0]	[0.6-0.7]	[4,390-4,682]	[130-139]	[22.4-24.0]
Italy	[110-136]	[3.2-4.0]	[0.6-0.7]	[4,230-4,539]	[125-134]	[21.5-23.3]
Spain	[116-137]	[3.4-4.0]	[0.6-0.8]	[4,429-4,652]	[131-137]	[22.5-24.0]
Germany	[126-150]	[3.6-4.3]	[0.7-0.8]	[4,311-4,655]	[127-138]	[22.0-23.9]

Table 7.14: Predicted usage pattern of antibiotics in the baseline case scenario and in the case of vaccination campaigns. The 95% RR of the daily number of administered antibiotics courses per 100,000 at its peak is reported, along with the total amount predicted to be administered by the end of the pandemic wave. Results are shown for several countries in the Northern Hemisphere, broken down for different stages of influenza-associated complications. Pneumonia stages I, II and III corresponds to home-treatment (or supervised outpatient treatment), hospital wards and ICU, respectively (see Figure (7.9) and Table (7.12)).

8

Conclusion

Dissertations are not finished;
they are abandoned.
F. Brooks

In this work we presented the framework of Reaction-Diffusion models on complex network topologies. Within this general approach different problems such as the importance of nodes in the WWW or the spreading of infectious diseases can be analyzed and discussed. We proposed a theoretical modelization of these processes giving new results and interpretations based on it. A description of the analytical treatment of each process has been given, as well as a detailed discussion of the numerical procedures which have been used.

In the first part of this thesis we tackle the problem of how to evaluate the importance of nodes in complex networks. One of the main feature of these systems is the absence of a characteristic scale in many topological properties. Not all the nodes are the same, to sort out their differences has become a relevant problem in different fields ranging from biology to data retrieval. We discuss this problem as a diffusive process in which the importance of a node is passed from one node to the other nodes that are linked to it. In this way one is able to take into account in a very efficient way the entire topology of a complex network and not just local, possibly misleading, quantities. We have studied the most important centrality measures based on properties of graph matrices: PageRank,

Eigenvector centrality, and the hub and authority scores of HITS. All these measures deduce the importance of a node in a self-consistent way from the importance of its nearest neighbors and, in the case of the HITS scores, of its next-to-nearest neighbors. Resuming some recent results on PageRank distributions on particular types of tree-like graphs we studied for the first time the extension of PageRank to the case of undirected networks, finding that the reduced PageRank of a node is proportional to its degree, for large degrees, for any graph and value of damping factor q. Similarly, the reduced α-centrality of a node is also proportional to its degree, for large degrees, on any graph. Within the same type of argument it is possible to show that the authority score of a node is proportional to its indegree, for large indegrees, when the outdegrees of all nodes are (approximately) the same. There are often strong relations between our centrality measured and (in)degree: some relations apply only to particular graphs and/or limits, others are more general. Our new findings imply that the measures are often strongly correlated with each other. We have indeed seen that the rankings of nodes according to the centrality measures we have considered are quite close to each other for indegree, PageRank, Eigenvector centrality and authority score on graphs built with the prescription of Dorogovtsev, Mendes and Samukhin. We have shown that these graphs have special properties, and that some measures may be correlated to each other. Instead, on real graphs, like the networks of political blogs and the sample of the Web graph we have considered, the structure is less regular and the measures are far less correlated to each other, as confirmed by the small values of the Kendall's τ for each pair of centrality measures. This means that, in spite of their similarities, spectral centrality measures look at nodes from different perspectives, thus giving different, and complementary, information about their importance. Also the scores computed from spectral centrality measures can complement the information about node's centrality derived from more traditional measures like node betweenness. This is especially important for directed graphs, where a variety of insights in interpreting network theory, for examples locating "repulsive" regions in a Web and relating hub formation to Anderson-like localizations, allowed us to define a new method to evaluate PageRank in a fast way.

The second part of this thesis has been devoted to the study of the spreading of infectious diseases. We presented a new theoretical approach, within a single population, to model behavioral changes that might be experienced during a severe outbreak. This

is still an open issue in epidemiology. We proposed a new model, *fear model*, to describe different mechanisms that people might use to react during a deadly epidemic disease. This part has been mainly a theoretical effort and it contains a detailed description of all the thrilling features such as: second peaks in the incidence curves and first order phase transitions related to endemic state of altered behaviors. The presence of a phase space in which two peaks are allowed is a very important feature not present in the basic epidemic models. In the historical data of the Spanish Flu of 1918 is clear how in many American cities two peaks has been experienced. Recently the presence of two peaks has been correlated to behavioral changes due by the severity of the epidemic. The endemic state of fear associated to a first order phase transition is an extremely important property of the model that clearly shows nontrivial dynamics in its phase space. We proved how it reduces the impact of a second epidemic on the populations.
Following this line of argument a general framework of metapopulation models has been introduced. In these models populations are nodes of a network. They represent cities with a local population connected with other cities through mobility of individuals. This multi-scale approaches explicitly include demographic and mobility heterogeneities. Detailed calculations of the global epidemic threshold which determines the invasion dynamics of the subpopulations are reported and the threshold critical dependence on the disease parameters and on the diffusion rates of the individuals is shown. We have also discussed the implications for disease extinction in the coupling between subpopulations. After the description of the basic results we introduced a new model in order to take into the account more realistic diffusion protocols. We analytically calculated, in two opposite limits, the global invasion threshold in the case of origin-destination diffusion processes. Our result turn out to be in very good agreement with the numerical simulations and opened a variety of future applications and generalizations.

In the third and last part of this work we moved further on with the introduction of a discrete stochastic epidemic computational model based on a metapopulation approach, the *Global Epidemic and Mobility* model, GLEaM. It is a data-driven model, in which the whole globe is partitioned in geographical census areas connected in a network of interactions by real data about human travel fluxes corresponding to transportation infrastructures and mobility patterns. To build this realistic model we have used the previously described advances in metapopulations theory. We applied GLEaM in the

recent pandemic. In particular we have defined a Monte Carlo likelihood analysis for the assessment of the seasonal transmission potential of the new A(H1N1) influenza. The analysis is based on the chronology of case detection in affected countries at the early stage of the epidemic. This method allowed the use of data coming from the border controls and the enhanced surveillance aimed at detecting the first cases reaching uninfected countries. This data is, in principle, more reliable than the raw count of cases provided by countries during the evolution of the epidemic and provided the necessary input to our large-scale computational model for the analysis of the future unfolding of the pandemic. Intriguingly our analysis showed the potential for an early activity peak that strongly emphasized the need for detailed planning for additional intervention measures, such as social distancing and antiviral drugs use, to delay the epidemic activity peak and thus increase the effectiveness of the subsequent vaccination effort. Our forecast have been proven, several months later, to be really accurate all-over the world. The observed peaks were in good agreement with the result coming out from our model: within the 95% reference range of the simulations. By using GLEaM it was possible to provide a model estimate of the number of imported cases arriving from Mexico to a set of selected countries. The dates and target countries were chosen to facilitate the comparison with the numbers found in the literature. We considered the importation of infected/exposed individual traveling from Mexico in one of the listed countries as of the date of May the 8^{th} 2009. For each country the numbers of imported cases are typically small, and as such prone to large stochastic fluctuations. However the surveillance values are all within the 99% reference ranges of the $2 \cdot 10^3$ realizations of our model.

We have also considered that in the US the travel history is known only for 50% of the confirmed cases. The simple extrapolation that provides a twofold estimate of imported cases is however still compatible with the reference range of our stochastic simulations. While GLEaM takes into account a higher level of geographical organization than previous approaches, its estimates still contain a number of assumptions and approximations. The contagion within each census area is approximated by means of a homogeneous mixing process. Once a person arrives at a census area by plane, he/she comes integrated into the local population. This implies that the travelers and the local population are equally exposed to the disease. The model considers each individual as independent and the possibility of cluster cases is not considered. Despite the use of different approximations, our and other different approaches provided support to the possibility of a

reporting ratio of infected cases in Mexico as low as 1 in 100, in agreement with prior estimates. This finding is important when evaluating the massive amount of data which are now being collected in a large number of countries around the world. We can easily imagine that the reporting rate as well as any estimate of the cumulative attack rate in most of the countries could be easily underestimated by orders of magnitude. We closed this work presenting a model for the critical care demand and antibiotics resources during the pandemic. We found, as was confirmed later, that even in the worst case scenario when no interventions are implemented the demand would be well below the national average capacity of most of the countries.

This work has been conceived to be a path into dynamical processes on complex topologies. Different subjects have been studied within the same holistic approach. Different techniques from several fields of Physics have been used. We analyzed problems that can be framed in computer sciences, social sciences and epidemiology facing them with random walk theory, localization and critical phenomena, multiscale analysis and ab intio models that are all well grounded in Physics. The used of these approaches turn out to be relevant and decisive to get a deep understanding on the processes analyzed. From very simple and abstract diffusion phenomena we moved into more realistic, coupled and multiscale dynamics. This shift has been progressive. All the pieces have been added one by one. Using the theory of random walk and localization phenomena we analyzed centrality diffusion of nodes and PageRank localization. Social disruption due to epidemic spreading has been studied from a theoretical point of view, considering realistic mechanisms that might induce individuals into a population to change their behaviors. The coupling terms we proposed create a nontrivial phase space in which critical phenomena can be observed. We shifted then our attention to multiscale systems in which populations are coupled due to the diffusion of individuals. We first introduced the general theory of metapopulation models on networks considering Markovian diffusion process showing and analyzing their critical behaviors. Then we moved into more realist protocols of diffusion considering origin and destination matrices. Using all these ingredients and multiscale analysis we proposed an ab initio, data driven and structured model for realistic forecasts of epidemic spreading in a world wide scale.
In all these processes the role of complexity in the diffusion has been point out and, when possible, analytically motivated. To consider complex features of networks is necessary

to understand the dynamics occurring on top of them.
Physics played a crucial role in the whole thesis giving us the tools to tackle and solve all the topics analyzed. It has been the necessary glue in this multidisciplinary work and hopefully showed how it can bring light not just in the mystery of the universe but even in problems that affect our daily life either during online searches or helping policy makers to provide for efficient plans of vaccination during the spreading of a infectious disease.
When Physics speaks, we must listen.

Acknowledgements

This work has been done between 2 continents in collaborations with people of 7 different countries, so I have to say thanks to a huge number of people with whom I have been honored to work. This time it was a bit more difficult to write down this page. So many people had an important role in this last chapter of the my education. Let me continue in my native language.

Il primo grazie è naturalmente rivolto a prof. Alex Vespignani. Per la sua pazienza, le sue indicazioni, le grandiose occasioni che mi ha concesso, le cose che mi ha insegnato e per essere sempre, oltre un grande esempio, una fonte enorme di ispirazione. Ha reso tutto questo possibile. Più di un grande boss, una vera e propria Guida.

Sono profondamente in debito anche verso prof. Gianni Mula e dott. Alessandro Chessa per avermi iniziato e fatto appassionare al mondo della *complessità*, per tutto il sostegno, tutti gli insegnamenti e ispirazioni.

Ringrazio prof. Guido Caldarelli per tutte le discussioni, le interazioni, le idee, la fiducia e le occasioni concessemi.

Ringrazio dott. Santo Fortunato per il tempo dedicatomi, tutte le interazioni e le idee discusse.

Ringrazio prof. Yamir Moreno e (quasi) dott. Sandro Meloni per tutte le interazioni, discussioni, piacevoli cene e risate che hanno reso grandiosa la nostra collaborazione

Ringrazio prof. Fil Menzer e prof. Alessandro Flammini per il tutto il

supporto, tutte le discussioni, l'infinita disponibilità e tutte le meravigliose partite a biliardino.

Thanks to Duygu for her patience, time, all the discussions, debugging, all the things she taught me, Turkish dishes, attemps to burn the house and her friendship.

Thanks to Bruno for all the crazy ideas we have dissussed, his time and patience, to have an answer to all my c++ questions and his friendship.

Thanks to Hao for his disponibility, patience, to have always a solution to crazy computational issues and for his friendship.

Ringrazio Al, Ben, Ross, Marcello, Francesco, Claudio e Giovà per l'amicizia, le folli serate e meriggi domenicali passati insieme a Bloomington, Indiana.

Ringrazio Paolo, Zap, Giamba, Scrillo, Fra, Amit e Berutti per essere stati amici e compagni spettacolari in questi 8 anni di deliri universitari.

Ringrazio Marco, Giovanni, Pier, Mirko, Manu e Edo per essere ormai dei fratelli. È difficile datare la nostra forte amicizia.

Ringrazio Ale, Giacco e Bise perché ormai sta iniziando a diventare difficile datare la nostra forte amicizia.

Ringrazio Fede per essere come una sorella.

Ringrazio Nicoletta per essere stata stupenda e per tutto il tempo passato insieme.

Ringrazio Nicoletta (Lai), Fra, Gio per tutto il tempo passato insieme e la forte amicizia.

Ringrazio Starbucks per avermi ormai reso dipendente di quella sostanza nera che chiamano espresso.

Ringrazio Delta, US, United, Alaska, Frontier, American airways, Continental, Alitalia, Meridiana e AirOne per avermi, nonostante tutto, portato in giro per il mondo.

Ringrazio Bloomington per essere una college town dove davvero tutto è possibile.

Ringrazio infine tutta la mia famiglia, per il supporto, affetto e la tranquillità trasmessami.

References

[1] Binney, J.J. and Dowrick, N.J. and Fisher, A.J. and Newman, M.E.J. *The Theory of Critical Phenomena*. Oxford Science Pubblications, 1992. 1

[2] Mandelbrot, B.B. *The Fractal Geometry of Nature*. Freeman, 1982. 1

[3] C.T. Butt. Revisiting the foundations of networks analysis. *Science*, 325:414–416, 2009. 1

[4] Bollobás, B. *Modern Graph Theory*. Springer-Verlag, 1998. 2, 10, 69

[5] Bollobás, B. *Random Graphs*. Cambridge studies in advanced mathematics, 1985. 2, 10

[6] P.W. Anderson. More is different. broken symmetry and the nature of the hierarchical structure of science. *Science*, 177, 1972. 2

[7] S. Milgram. The small world problem. *Psychoology Today*, **2**:60–67, 1967. 3, 21

[8] P. Erdös and A Rényi. *Publ. Math.*, **6**:290, 1959. 3, 22

[9] P. Erdös and A Rényi. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**:17, 1960. 3, 22

[10] P. Erdös and A Rényi. *Bull. Inst. Int. Stat.*, **38**:343, 1961. 3, 22

[11] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, **393**:440–442, 1998. 3, 20, 24

[12] A.-L. Barabási and R. Albert. *Nature*, **286**:509, 1999. 3, 20, 21, 26

[13] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, **30**:107–117, 1998. 4, 59

[14] Murray, J.D. *Mathematical Biology.* 3rd edition Berlin: Springer Verla, 2005. 4, 112

[15] J.S.M. Peiris, K.Y. Yuen, and K. Stohr. *N. Engl. J. Med.*, 349:2431, 2003. 4

[16] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J.J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, and A. Vespignani. Seasonal transmission potential and activity peaks of the new influenza a(h1n1): a monte carlo likelihood analysis based on human mobility. *BMC Medicine*, **439**:7:45, 2009. 4, 89, 107, 150, 161, 162, 163, 164

[17] A. Vespignani. Predicting the behavior of techno-social systems. *Science*, 325:425–428, 2009. 6, 89

[18] R.M. Anderson and R.M. May. Spatial, temporal and genetic heterogeneity in hosts populations and the design of immunization programs. *IMA J. Math. Appl. Med. Biol.*, 1:233–266, 1984. 6, 106

[19] B.M. Bolker and Grenfell. B.T. Space persistence and dynamics of measles epidemics. *Phil. Trans. Biol. Sci.*, 348:309–320, 1995. 6, 106

[20] A.L. Lloyd and R.M. May. Spatial heterogeneity in epidemic models. *J. Theor. Biol.*, 179:1–11, 1996. 6, 106

[21] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J.J. Ramasco, and Vespignani A. Multiscale mobility networks and the large scale spreading of infectious diseases. *Proc. Natl Acad. Sci.*, page 106:21484, 2009. 6, 7, 44, 89, 106, 107, 125, 128, 130, 143, 168

[22] V. Colizza and A. Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.*, 99:148701, 2007. 6

[23] V. Colizza and A. Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of Theoretical Biology*, 251:450–467, 2008. 6, 115

[24] Barrat, A. and Barthèlemy, M. and Vespignani, A. *Dynamical Processes on Complex Networks.* Cambridge University Press, 2008. 9, 32, 38, 47, 50

[25] Newman, M.E.J. *Networks, an Introduction.* Oxford University Press, 2010. 9

[26] Bergé, C. *Graphs and Hypergraphs.* North-Holland, 1976. 10

[27] Chartrand, G. and Lesniak, L. *Graphs and Digraphs.* Wadsworth and Brooks/Cole, 1986. 10

[28] Clark, J. and Holton, D.A. *A First Look at Graph Theory.* World Scientific, 1991. 10

[29] Harary, F. *Graph Theory.* Perseus, 1995. 10

[30] West, G.B. *Introduction to Graph Theory.* Prentice Hall, 1996. 10

[31] M. Granovetter. Strength of weak ties. *American Journal of Sociology*, **78**:1360–1380, 1973. 11

[32] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of Internet: A Statistical Physics Approach.* Cambridge University Press., 2004. 11, 20, 44

[33] A. Barrat, M. Barthélémy, R. Pastor-Satorras, and A. Vespignani. The architeture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**:3747–3752, 2004a. 11, 18, 20, 108, 114, 128

[34] D Garlaschelli, S. Battiston, S. Castri, M. Servedio, and G. Caldarelli. The scale free topology of market investments. *Physica A.*, **350**:491–499, 2005. 14

[35] L.C. Freeman. A set of measures of centrality based upon betweenness. *Sociometry*, **13**:141–154, 1977. 15

[36] U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25:163–177, 2001. 15

[37] M. Boguña and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, **66**:047104, 2002. 17

[38] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, **89**:208701, 2002a. 17

[39] Moreno, J.L. *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodram.* Beacon House, 1934. 19

[40] H. Ebel, L.I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, **66**:035103, 2002. 19

[41] M. E. J. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, **66**:035101, 2002c. 19

[42] J.J. Ramasco, S.N. Dorogovtsev, and R. Pastor-Satorras. Self-organization of collaboration networks. *Phys. Rev. E*, **70**:036106, 2004. 20

[43] M.E.J. Newman, S.H. Strogatz, and D.J. Watts. *Physical Review E*, **64**:026118, 2001. 20

[44] M.E.J. Newman. Scientific collaboration networks. I. Network contruction and fundamental results. *Phys. Rev. E*, **64**:016131, 2001. 20

[45] M.E.J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks and centrality. *Phys. Rev. E*, **64**:016132, 2001. 20

[46] M.E.J. Newman. The structure of scientific collaboration networks. *Proc. Natl Acad. Sci.*, **98**:404–409, 2001. 20

[47] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Staneeley. Classes of small-world networks. *Proceeding of the National Academy of Science (USA)*, **97**:11149–11152, 2000. 20

[48] A. De Montis, M. Barthélémy, A. Chessa, and A. Vespignani. The structure of inter-urban traffic: A weighted network analysis. *Env. Planning Journal B*, 2006. 20

[49] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *Computer Communications Review*, **29**:251–262, 1999. 20

[50] Q. Chen, H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger. The origin of power laws in Internet topologies revisited. *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies*, 2002. 20

[51] R. Guimerá and L.A.N. Amaral. Modeling the world-wide airport network. *Eur. Phys. J.B.*, **38**:381–385, 2004. 20

[52] R. Albert, H. Jeong, and A.-L. Barabási. Internet: Diameter of the world wide web. *Nature*, **401**:130–131, 1999. 20

[53] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. *Computer Networks*, **33**:309–320, 2000. 20

[54] U. Alon. Biological networks: the tinkerer as an engineer. *Science*, **301**:1866–1867, 2003. 20

[55] A.-L. Barabási and Z.N. Oltvai. Network biology: understading the cell's functional organization. *Nat. Rev. Gen.*, **5**:101–113, 2004. 20

[56] H. Jeong, S. Mason, A.-L. Barábasi, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, **411**:41–42, 2001. 20

[57] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, and Y. Aberg. The web of human sexual contacts. *Nature*, **411**:907–908, 2001. 20

[58] J.A. Dunne, R.J. Williams, and N.D. Martinez. Food-web structure and network theory: the role of connectance and size. *Proc. Natl Acad. Sci.*, **99**:12917–12922, 2002. 20

[59] F. Chung and .H Lu. *Advances Applied Mathematics*, **26**:257, 2001. 23

[60] A. Barrat and M. Weigt. *Eur. Phys. J. B*, **13**:547, 2000. 24

[61] A.-L. Barabási, R. Albert, and H. Jeong. *Physica A*, **272**:173, 1999. 26

[62] S.N. Dorogovtsev, J.F.F Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letter*, **85**:4633–4636, 2000. 26, 28, 64

[63] P.L. Krapivsky, S. Redner, and F. Leyvraz. *Physical Review Letter*, **85**:4629, 2000. 26

[64] B. Bollobás and O. Riordan. The diameter of scale-free random graphs. 2002. 27

[65] P.L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, **63**:066123, 2001. 28

[66] C.W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and Natural Sciences*. Springer-Verlag, 1990. 32

[67] Ma, S.K. *Statistical Mechanics*. World Scientific, 1985. 32

[68] Chandler, D. *Introduction to Modern Statistical Mechanics*. Oxford University Press, London (UK), 1987. 32

[69] Balescu, R. *Statistical Dynamics: Matter Out of Equilibrium*. Wiley, 1997. 32

[70] Huang, K. *Statistical Mechanics*. Wiley, 1987. 32

[71] Economou, E.N. *Green's functions in quantum physics*. Springer, 2006. 38

[72] G.J. Rodgers and A.J. Bray. Density of states of a sparse random matrix. *Phys. Rev. B.*, 37:3557–3562, 1988. 38

[73] R. Monasson. Diffusion, localization and dispersion relations on small-world. *Eur. Phys. J. B*, 12:555–567, 1999. 39

[74] A. Samukhin, S. Dorogovtsev, and J. Mendes. Laplacian spectra and random walks on complex networks: are scale-free architectures really important? *Phys. Rev. E*, 77:036115, 2008. 39

[75] J.D. Noh and H. Rieger. Random walks on complex networks. *Phys. Rev. Lett.*, 92:118701, 2004. 39

[76] Anderson, R.M. and May, R.M. *Infectious Diseases in Humans*. Oxford Univ. Press, 1992. 39, 44, 106, 146

[77] O. Diekmann, A.J.P. Heesterbeek, and J.A.J. Metz. On the definition and the computation of the basic reproduction ratio r0 in models for infectious diseases in heterogeneous populations. *J. Math. Bio.*, 28:1432, 1990. 39

[78] Keeling, M.J. and Rohani, P. *Modeling infectious diseases in humans and animals*. Pronceton Univeristy Press, 2008. 40

[79] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *P Roy Soc Long A Mat.*, **115**:700–721, 1927. 41, 90

[80] Bailey, N.T. *The mathematical theory of infectious diseases.* Griffin, 1975. 43, 112

[81] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001. 44, 47

[82] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. Epidemic spreading in complex networks with degree correlations. *Lect. Notes Phys.*, 625:127–147, 2004. 45

[83] M. Boguñá and R. Pastor-Satorras. Class of correlated random networks with hidden variables. *Phys. Rev. E*, 68:036112, 2003. 46

[84] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B*, 26:521–529, 2004. 49

[85] R.M. May and A.L. Lloyd. Infectious dynamics on scale-free networks. *Phys. Rev. E*, 64:066112, 2001. 49

[86] R. Pastor-Satorras and A. Vespignani. Immunization of complex networks. *Phys. Rev. E*, 65:035108, 2002. 50

[87] P. Bonacich and P. Lloyd. *Soc. Netw.*, 23:191, 2001. 60, 62

[88] J.M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**:604–632, 1999. 60, 62

[89] P. Boldi, M. Santini, and S. Vigna. Pagerank as a function of the damping factor. *WWW2005*, pages 557–566, 2005. 61, 83

[90] S. Fortunato, A. Flammini, M. Boguñá, and F. Menczer. How to make the top ten: Approximating PageRank from in-degree. *arXiv.org,physics,physics/0511103*, 2005. 61

[91] S. Fortunato and A. Flammini. Random Walks on Direct Networks: the Case of PageRank. *Int, J. Bif. Ch.*, 17, 2007. 61, 64, 67, 83

[92] A.-L. Barabási and R. Albert. *Science*, **286**:509, 1999. 64, 67

[93] J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web as a graph: measurements, models and methods. *LNCS*, **1627**:1–18, 1999. 67

[94] P. De Los Rios. *Europhys. Lett.*, 56:898, 2001. 67

[95] G. Caldarelli, R. Marchetti, and L. Pietronero. *Europhys. Lett.*, 52:386, 2000. 67

[96] P. Chen, H. Xie, S. Maslov, and S. Redner. *J. Informet.*, 1:8, 2007. 68

[97] M. Kendall. *Biometrika*, 30:81, 1938. 73

[98] L.A. Adamic and N. Glance. The political blogosphere and the 2004 us election. *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005. 74

[99] N. Perra, V. Zlatić, A. Chessa, C. Conti, D. Donato, and G. Caldarelli. Pagerank equation and localization in the www. *EPL*, 88:48002, 2009. 81

[100] Langville, A.N. and Meyer, C.D. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, 2006. 81

[101] D. Garlaschelli and M.I Loffredo. Patterns of link reciprocity in directed networks. *Phys. Rev. L*, **93**:188701, 2004. 83

[102] Stauffer, D. and Aharony, A. *Introduction to Percolation Theory.* Taylor and Francis, 1994. 83

[103] P. Boldi and S. Vigna. The web graph framework i: Compression techniques. *WWW2004*, pages 595–601, 2004. 84

[104] Sommerfeld A. *Partial Differential Equations in Physics.* Academic Press, New York, 1949. 87

[105] Zinn-Justin J. *Quantum Field Theory and Critical Phenomena.* Oxford University Press, London (UK), 2002. 87

[106] N.M. Ferguson. Capturing human behaviour. *Nature*, page 466:733, 2007. 89

[107] J. Wallinga, P. Teunis, and M. Kretzschmar. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am. J. Epi.*, **164**, 2006. 89

[108] D. Brockmann, L. Hufnagel, and L. Geisel. The scaling laws of human travel. *Nature*, **439**:462–465, 2006. 89

[109] F. Ginelli, H. Hinrichsen, R. Livi, D. Mukamel, and A. Torcini. Contact processes with long-range interactions. *J. Stat. Mech.*, 2006. 89

[110] P. Bajardi, C. Poletto, D. Balcan, H. Hu, B. Goncalves, J.J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, V. Colizza, and A. Vespignani. Modeling vaccination campaigns and the fall/winter 2009 activity of the new a(h1n1) influenza in the northern hemisphere. *Emerging Health Threats*, **2:e11**, 2009. 89, 153, 163, 168

[111] Gardner, D. *The Science of Fear: How the Culture of Fear Manipulates Your Brain.* Plume, 2009. 89

[112] R.J. Hatchett, C.E. Mecher, and Lipsitch M. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc. Natl Acad. Sci.*, **104**:7582–7587, 2007. 89, 90

[113] G. Cruz-Pacheco, L. Duran, L. Esteva, A.A. Minzoni, M. López-Cervantes, P. Panayotaros, A. Ahued, and I. Villaseñor. Modelling of the influenza a(h1n1)v outbreak in mexico city, april-may 2009, with control sanitary. *EuroSurveillance*, 14:19254, 2009. 89, 144, 147, 162

[114] M.C.J. Bootsma and N.M. Ferguson. Public health interventions and epidemic intensity during the 1918 influenza pandemic. *Proc. Natl Acad. Sci.*, **104**:7588–7593, 2007. 90, 96

[115] H. Markel, H.B. Lipman, J.A. Navarro, A. Sloan, J.R. Michalsen, A.M. Stern, and Cetron M.S. Nonpharmaceutical interventions implemented by us cities during the 1918-1919 influenza pandemic. *JAMA*, **298**:6, 2007. 90

[116] P. Poletti, B. Caprile, M. Ajelli, A. Pugliese, and Merler S. Spontaneous behavioural changes in response to epidemics. *J. Theor. Biol.*, pages 225–228, 2009. 90

[117] W. Goffman and V.A. Newill. Generalization of epidemic theory an application to the transmission of ideas. *Nature*, **4955**:225–228, 1964. 90

[118] J.M. Epstein, J. Parker, D. Cummings, and Hammond R.A. Coupled contagion dynamics of fear and disease mathematical and computational explorations. *PloS ONE*, **3**:E3955, 2008. 90

[119] S. Funk, E. Gilad, C. Watkins, and V.A.A. Jansen. The spread of awareness and its impact on epidemic outbreaks. *Proc. Natl Acad. Sci.*, 2009. 90

[120] Lynch A. Thought contagion as asbract evolution. *J Ideas*, **2**:3–10, 1991. 90

[121] DeFleur, M.L. and Ball-Rokeach, S. *Theories of Mass Communication*. Longman, NY, 1989. 90

[122] N.M. Ferguson, M.J. Keeling, and W.J. et al Edmunds. Planning for smallpox outbreaks. *Nature*, 425:681–685, 2003. 106

[123] D.J. Watts, R. Muhamad, D.C. Medina, and P.S. Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulations model. *Proc. Natl. Acad. Sci.*, 102:11157–11162, 2005. 106

[124] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. The modeling of global epidemics: Stochastic dynamics and predictability. *Bull Math Biol*, **68**:1893–1921, 2006. 106, 128

[125] D.J.D. Earn, P. Rohani, and B.T. Grenfell. Persistence, chaos and synchrony in ecology and epidemiology. *Proc. Roy. Soc. Lond. B*, 265:7–10, 1998. 106

[126] P. Rohani, D.J.D. Earn, and B.T. Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286:968–971, 1999. 106

[127] M.J. Keeling. Metapopulation moments: coupling, stochasticity and persistence. *Journal of Animal Ecology*, 69:725–736, 2000. 106

[128] A.W. Park, S. Gubbins, and C.A. Gilligan. Extinction times for closed epidemics: the effects of host spatial structure. *Ecology Letters*, 5:747–755, 2002. 106

[129] C. Viboud, O. N. Bjornstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312:447, 2006. 106

[130] O.V. Baroyan, L.A. Genchikov, L.A. Rvachev, and V.A. Shashkov. An attempt at large-scale influenza epidemic modelling by means of a computer. *Bull. Int. Epidemiol. Assoc.*, 18:22–31, 1969. 107

[131] L.A. Rvachev and I.M. Longini. A mathematical model for the global spread of influenza. *Mathematical Biosciences*, 75:3–22, 1985. 107, 150

[132] I.M. Longini. A mathematical model for predicting the geographic spread of new infectious agents. *Math. Biosci*, 90:367–383, 1988. 107

[133] R.F. Grais, H.J. Ellis, and G.E. Glass. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *Eur. J. Epidemiol.*, 18:1065â1072, 2003. 107

[134] R.F. Grais, J.H. Ellis, A. Kress, and G.E. Glass. Modeling the spread of annual influenza epidemics in the u.s.: The potential role of air travel. *Health Care Manag Sci*, 7:127, 2004. 107, 139

[135] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci.*, 101:15124, 2004. 107

[136] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS*, 103:2015, 2006. 107

[137] F. Ball, D. Mollison, and G. Scalia-Tomba. Epidemics with two levels of mixing. *Ann. Appl. Probab.*, 7:46–89, 1997. 110, 111

[138] P. Cross, J.O. Lloyd-Smith, P.L.F. Johnson, and M.G. Wayne. Duelling timescales of host movement and disease recovery determine invasion of disease in structured populations. *Ecol. Lett.*, 8:587–595, 2005. 110

[139] Harris, T.E. *The theory of branching processes.* Dover Publications, 1989. 111

[140] A. Vázquez. Epidemic outbreaks on structured populations. *J. Theor. Biol.*, 245:125–129, 2007. 111

[141] N.M. Ferguson, D.A.T. Cummings, C. Fraser, J.C. Cajka, P.C Cooley, and D.S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442:448–452, 2006. 126, 144

[142] C. Germann, K. Kadau, I.M. Longini, and C.A. Macken. Mitigation strategies for for pandemic influenza in the united states. *Proc. Nat. Acad. Sci.*, 103:5935–5940, 2006. 126, 144

[143] M.E. Halloran, N.M. Ferguson, S. Eubank, I.M. Longini, D.A.T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T.C. Germann, D. Wagener, R. Beckman, K. Kadau, C.A. Macken, D.S. Burke, and P. Cooley. Modeling targeted layered containment of an influenza pandemic. *Proc. Natl. Acad. Sci*, 105:4639–4644, 2008. 126

[144] M.L. Ciofi degli Atti, S. Merler, C. Rizzo, M. Ajelli, and M. Massari et al. Mitigation measures for pandemic influenza in italy: An individual based model considering different scenarios. *PLoS ONE*, 3:e1790, 2008. 126

[145] N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437:209, 2005. 126, 156

[146] S. Merler and M. Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proc. Royal Soc. B: Biological Sciences*, 277:557–565, 2010. 126

[147] Center for international earth science information network (ciesin), columbia university; and centro internacional de agricultura tropical (ciat). the gridded population of the world version 3 (gpwv3): Population grids. palisades, ny: Socioeconomic data and applications center (sedac), columbia university. 126

[148] Center for international earth science information network (ciesin), columbia university; international food policy research institute (ifpri); the world bank; and centro internacional de agricultura tropical (ciat). global rural-urban mapping project (grump), alpha version: Population grids. palisades, ny: Socioeconomic data and applications center (sedac), columbia university. 126

[149] Okabe, A. and Boots, B. and Sugihara, K. and Chiu, S.N. *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams*. John Wiley, 2000. 127

[150] V. Colizza, A. Barrat, M. Barthelemy, and A. Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci.*, **103**:2015–2020, 2006. 128

[151] V. Colizza, A. Barrat, M. Barthelemy, A.J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Med*, **4**, 2007. 128, 144, 149, 157

[152] L. Sattenspiel and K. Dietz. A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, **128**:71–91, 1995. 135

[153] M.J. Keeling and P. Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, **5**:20–29, 2002. 136

[154] B.S. Cooper, R.J. Pitman, W.J. Edmunds, and N.J. Gay. Delaying the international spread of pandemic influenza. *PLoS Medicine*, 3:e212, 2006. 139, 147, 158

[155] D. Balcan, B. Goncalves, H. Hu, J.J. Ramasco, V. Colizza, and Vespignani A. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science*, 1:3:132–145, 2010. 143

[156] S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004. 144

[157] C. et al Fraser. Pandemic potential of a strain of influenza a(h1n1): early findings. *Science*, 324:1557–1561, 2009. 144, 146, 149, 161, 163

[158] Who, pandemic (h1n1) 2009 briefing note 3 (revised): Changes in reporting requirements for pandemic (h1n1) 2009 virus infection. 144, 152

[159] K. Khan, J. Arino, W. Hu, P. Raposo, J. Sears, F. Calderon, C. Heidebrecht, M. Macdonald, J. Liauw, A. Chan, and M. Gardam. Spread of a novel influenza

a(h1n1) virus via global airline transportation. *N Engl J Med*, 361:212–214, 2009. 144

[160] I.M. Longini, M.E. Halloran, A. Nizam, and Y. Yang. Containing pandemic influenza with antiviral agents. *Am J Epidemiol*, 159:623, 2004. 146, 149, 156, 157

[161] F. Carrat, E. Vergu, N.M. Ferguson, M. Lemaitre, S. Cauchemez, S. Leach, and A.J. Valleron. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am J Epidemiol*, **167**:775–785, 2008. 146, 149

[162] I.M. Longini, A. Nizam, Xu S., K. Ungchusak, W. Hanshaoworakul, D.A.T. Cummings, and M.E. Halloran. Containing pandemic influenza at the source. *Science*, 309:1083–1087, 2005. 146, 149, 156, 157

[163] Brote de infeccion respiratoria aguda en la gloria, municipio de perote, mexico secretaria de salud, mexico. 146, 147

[164] Who wkly epidemiol rec, 2009. 148

[165] Cdc interim guidance for clinicians on identifying and caring for patients with swine-origin influenza a (h1n1) virus infection, 2009. 148

[166] F.S. Dawood, S. Jain, L. Finelli, M.W. Shaw, and S. et al Lindstrom. Novel swine-origin influenza a (h1n1) virus investigation: Emergence of a novel swine-origin influenza a (h1n1) virus in humans. *N Engl J Med*, 360:2605–2615, 2009. 148

[167] M.J. Roberts and J.A.P. Heesterbeek. Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. *J. Math. Bio.*, 55:803–816, 2007. 149

[168] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc R Soc B*, **274**:599–604, 2007. 149

[169] R. Gani, C. H. Hughes, D. Fleming, T. Griffin, J. Medlock, and S. Leach. Potential impact of antiviral drug use during influenza pandemic. *Emer. Inf. Dis.*, 11:1355, 2005. 149, 156

[170] L.R. Elveback, J.P. Fox, E. Ackerman, A. Langworthy, M. Boyd, and L. Gatewood. An influenza simulation model for immunization studies. *Am J Epidemiol*, 103:152–165, 1976. 149

[171] P.Y. Boelle, P. Bernillon, and J.C. Desenclos. A preliminary estimation of the reproduction ratio for new influenza a(h1n1) from the outbreak in mexico, march-april 2009. *EuroSurveilliance*, 14, 2009. 149, 150

[172] H. Nishiura, C. Castillo-Chavez, M. Safan, and G. Chowell. Transmission potential of the new influenza a(h1n1) virus and its age-specificity in japan. *EuroSurveilliance*, 14, 2009. 149, 150

[173] H. Nishiura, N.M. Wilson, and M.G. Baker. Estimating the reproduction number of the novel influenza a virus (h1n1) in a southern hemisphere setting: preliminary estimate in new zealand. *NZ Med J*, 122:1–5, 2009. 149, 150

[174] J. Lessler, N.G. Reich, R. Brookmeyer, T.M. Perl, K.E. Nelson, and D.A. Cummings. Incubation periods of acute respiratory viral infections: a systematic review. *Lancet Infect DisA*, 9:291–300, 2009. 149

[175] Cdc: Briefing on public health investigation of human cases of swine influenza, april 23, 2009, 3:30 p.m. es. 150

[176] Public health agency of canada, cases of h1n1 flu virus in canada, june 10 2009. 150

[177] King's-edgehill school web site, first cases in canada was related to a school trip in mexico. 150

[178] Who, chronology of influenza a(h1n1). 150

[179] El salvador journal,primeros casos de gripe porcina el salvador, may 4 2009. 150

[180] Ministerio de salud pública y asistencia social, 8th official update, may 3, 2009. 150

[181] Abc news, swine flu cases confirmed in scotland, april 28. 150

[182] The scottish government, scottish government news release, april 26. 150

[183] Ministerio de sanidad y political social, official update, april 27 2009. 150

[184] The guardian, spain confirms first swine flu case in europe, april 27 2009. 150

[185] Usa today, cuba confirms its 1st swine flu case, may 12 2009. 150

[186] Global post, costa rica reports first swine flu case, april 28 2009. 150

[187] Ministry of health welfare and sport (netherlands), first victim mexican flu, april 30 2009. 150

[188] Robert koch institut, neue influenza a/h1n1 in deutschland bewertung des bisherigen geschehens. 150

[189] Ministère de la santé et des sports, official update, may 1. 150

[190] El periodico guatemala, ministro de salud confirma primer caso de ah1n1 en guatemala, may 5 2009. 150

[191] Ministerio de salud pública de guatemala, official update, may 5 2009. 150

[192] Ministerio de la protecciìon social república de colombia, official update, may 3 2009. 150

[193] C.E. Mills, J.M. Robins, and M. Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432:904–906, 2004. 150

[194] Who official data. 153

[195] N. Wilson and M.G. Baker. The emerging influenza pandemic: estimating the case fatality ratio. *Euro Surveilliance*, 14:26, 2009. 153, 156

[196] Graske et al. Assessing the severity of the novel a/h1n1 pandemic. *BMJ*, 339:b2840, 2009. 156

[197] A. Flahault, E. Vergu, L. Coudeville, and R. Grais. Strategies for containing a global influenza pandemic. *Vaccine*, 24:6751–6755, 2006. 156

[198] T.C. Germann, K. Kadau, I.M. Longini, and C.A. Macken. Mitigation strategies for pandemic influenza in the united states. *Proc. Natl. Acad. Sci.*, 103:5935–5940, 2006. 156

[199] N. Arinaminpathy and A.R. McLean. Antiviral treatment for the control of pandemic influenza: some logistical constraints. *J. R. Soc. Interface*, 5:5945–553, 2008. 156

[200] J.T. Wu, S. Riley, C. Fraser, and G.M. Leung. Reducing the impact of the next influenza pandemic using household-based public health interventions. *PLoS Med*, 3:e361, 2006. 156

[201] Roche h-l: Update on current developments around tamiflu (2007). 156

[202] Singer et al. Meeting report: Risk assessment of tamiflu use under pandemic conditions. *Environ Health Perspect*, 116:1563â1567, 2008. 156

[203] M. Lipsitch, M. La jous, J.J. OHagan, T. Cohen, and J.C. Miller. *PLoS ONE*, 4:e6895, 2009. 161, 162, 163

[204] Eurosurveilliance. 161, 162

[205] Eurosurveilliance. 161, 162

[206] Reports of the brazilian health department (ministerio da saude). 161, 162

[207] Secretaria de salud, mexico. situation actual de la epidemia, oct 12, 2009. 163

[208] Uk department of health. swine flu: Uk planning assumptions. issued 3 september, 2009. 163, 168

[209] Reed C, Angulo FJ, Swerdlow DL, Lipsitch M, Meltzer MI, Jernigan D, and et al. Estimates of the prevalence of pandemic (h1n1) 2009, united states, aprilâjuly 2009, 2009. 163, 168

[210] European centre for disease control and prevention, pandemic (h1n1) 2009 daily update (november 23, 2009). 163, 168

[211] Morens DM, Taubenberger, and Fauci AS. Predominant role of bacterial pneumonia as a cause of death in pandemic influenza: implications for pandemic influenza preparedness. *J Infect Dis*, 198:962–970, 2008. 163

[212] Perez-Padilla R, de la Rosa-Zamboni D, Ponce de Leon S, and et al. Pneumonia and respiratory failure from swine-origin influenza a (h1n1) in mexico. *New Engl J Med*, 361:680–689, 2009. 163, 167

[213] CDC. Hospitalized patients with novel influenza a (h1n1) virus infection - california, april-may, 2009. *MMWR*, 58:536–541, 2009. 163

[214] CDC. Intensive-care patients with severe novel influenza a (h1n1) virus infection - michigan, june 2009. *MMWR*, 58:749–752, 2009. 163

[215] J. Rello, A. Rodríguez, P. Ibañez, and L. et al Socias. Intensive care adult patients with severe respiratory failure caused by influenza a (h1n1)v in spain, 2009. 163, 167, 168

[216] CDC. Bacterial coinfections in lung tissue specimens from fatal cases of 2009 pandemic influenza a (h1n1) - united states, may-august 2009. *MMWR*, 58:1–4, 2009. 163

[217] The ANZIC Influenza Investigators. Critical care services and 2009 h1n1 influenza in australia and new zealand. *New Engl J Med*, 361:1925–1934, 2009. 163, 165, 166, 167, 168, 169

[218] World health organization. clinical management of human infection with pandemic (h1n1) 2009: revised guidance, november 2009. 164

[219] W.S. Lim. Pandemic flu: clinical management of patients with an influenza-like illness during an influenza pandemic. *Thorax*, 62:1–46, 2007. 164, 165, 167, 168

[220] Society of critical care medicine, critical care statistics in the united states 2006. 164

[221] The information system of the federal health monitoring. 164

[222] Wunsch H, Angus DC, Harrison DA, and et al. Variation in critical care services across north america and western europe. *Crit Care Med*, 36(10):2787, 2008. 165

[223] Ercole A, Taylor BL, Rhodes A, and Menon DK. Modelling the impact of an influenza a/h1n1 pandemic on critical care demand from early pathogenicity data: the case for sentinel reporting. *Anaesthesia*, 64:937–941, 2009. 165

[224] British thoracic society. 165, 166, 167